US005768603A

United States Patent [19]

Brown et al.

[11] **Patent Number: 5,768,603**

[45] **Date of Patent: *Jun. 16, 1998**

[54] **METHOD AND SYSTEM FOR NATURAL LANGUAGE TRANSLATION**

[75] Inventors: **Peter Fitzhugh Brown**, New York; **John Cocke**, Bedford; **Stephen Andrew Della Pietra**, Pearl River; **Vincent Joseph Della Pietra**, Blauvelt; **Frederick Jelinek**, Briarcliff Manor; **Jennifer Ceil Lai**, Garrison; **Robert Leroy Mercer**, Yorktown Heights, all of N.Y.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[*] Notice: The term of this patent shall not extend beyond the expiration date of Pat. No. 5,477,451.

[21] Appl. No.: **459,698**

[22] Filed: **Jun. 2, 1995**

Related U.S. Application Data

[63] Continuation of Ser. No. 736,278, Jul. 25, 1991, Pat. No. 5,477,451.

[51] **Int. Cl.^{6}** **G06F 17/28**; G06F 17/20

[52] **U.S. Cl.** **395/759**; 395/751; 395/752; 395/757; 395/2.41; 395/2.48; 395/2.65; 395/2.86

[58] **Field of Search** 395/759, 2.65, 395/2.86, 751, 752, 757, 2.41, 2.48

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,754,489	6/1988	Bukser .
4,852,173	7/1989	Bahl et al. .
4,879,580	11/1989	Church .

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

327266	8/1989	European Pat. Off. .
357344	3/1990	European Pat. Off. .
399533	11/1990	European Pat. Off. .
WO 90/10911	9/1990	WIPO .

OTHER PUBLICATIONS

P. Brown, "Word-Sense Disambiguation Using Statistical Methods", *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Jun. 1991, pp. 264–270.

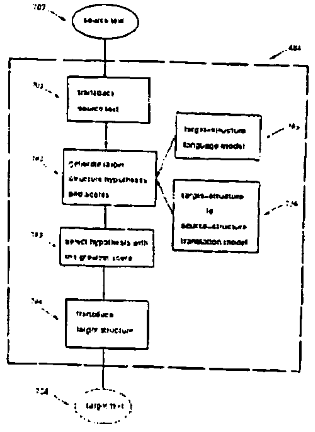
(List continued on next page.)

Primary Examiner—Gail O. Hayes
Assistant Examiner—Joseph Thomas
Attorney, Agent, or Firm—Sterne, Kessler, Goldstein & Fox, PLLC; Robert P. Tassinari

[57] **ABSTRACT**

The present invention is a system for translating text from a first source language into a second target language. The system assigns probabilities or scores to various target-language translations and then displays or makes otherwise available the highest scoring translations. The source text is first transduced into one or more intermediate structural representations. From these intermediate source structures a set of intermediate target-structure hypotheses is generated. These hypotheses are scored by two different models: a language model which assigns a probability or score to an intermediate target structure, and a translation model which assigns a probability or score to the event that an intermediate target structure is translated into an intermediate source structure. Scores from the translation model and language model are combined into a combined score for each intermediate target-structure hypothesis. Finally, a set of target-text hypotheses is produced by transducing the highest scoring target-structure hypotheses into portions of text in the target language. The system can either run in batch mode, in which case it translates source-language text into a target language without human assistance, or it can function as an aid to a human translator. When functioning as an aid to a human translator, the human may simply select from the various translation hypotheses provided by the system, or he may optionally provide hints or constraints on how to perform one or more of the stages of source transduction, hypothesis generation and target transduction.

34 Claims, 54 Drawing Sheets


source text
404
transduce source text
target-structure language model
generate target structure hypotheses and scores
target-structure to source-structure translation model
select hypothesis with the greatest score
transduce target structure
target text

U.S. PATENT DOCUMENTS

4,882,759	11/1989	Bahl et al. .	
4,984,178	1/1991	Hemphill et al. .	
4,991,094	2/1991	Fagen et al. .	
5,033,087	7/1991	Bahl et al. .	
5,068,789	11/1991	Van Vliembergen .	
5,072,452	12/1991	Brown et al.	395/2.65
5,109,509	4/1992	Katayama et al. .	
5,146,405	9/1992	Church	395/759
5,200,893	4/1993	Ozawa et al. .	
5,293,584	3/1994	Brown et al.	395/2.86
5,428,772	6/1995	Merz	395/604
5,444,617	8/1995	Merialdo	395/759
5,477,451	12/1995	Brown et al.	395/759

OTHER PUBLICATIONS

P. Brown et al., "Aligning Sentences in Parallel Corpora", *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Jun. 1991, pp. 169–176.

B. Merialdo, "Tagging Text With A Probalistic Model", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Paris, France, May 14–17, 1991.

M. Kay, "Making Connections", *ACH/ALLC '91*, Tempe, Arizona, 1991, p. 1.

"Method For Inferring Lexical Associations From Textual Co-Occurrences", IBM Technical Disclosure Bulletin, vol. 33, Jun. 1990.

L. Bahl et al., "A Tree-Based Statistical Language Model For Natural Language Speech Recognition", *IEEE Transactions of Acoustics*, vol. 37, No. 7, Jul. 1989, pp. 1001–1008.

J. Spohrer et al., "Partial Traceback in Continuous Speech Recognition", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, 1982.

F. Jelinek, R. Mercer, "Interpolated Estimated of Markov Source Parameters From Sparse Data", *Workshop on Pattern Recognition in Practice*, Amsterdam (Netherland), North Holland, May 21–23, 1980.

J. Baker, "Trainable Grammars For Speech Recognition", *Speech Communications Papers Presented at the 97th Meeting of the Acoustic Society of America*, 1979, pp. 547–550.

M.E. Lesk, "Lex–A Lexical Analyzer Generator", *Computer Science Technical Report*, No. 39, Bell Laboratories, Oct. 1975.

L. Baum, "An Inequality and Associated Maximation Technique in Statistical Estimation for Probalistic Functions of Markov Processes", *Inequalities*, vol. 3, 1972, pp. 1–8.

F. Jelinek, "Self Organized Language Modeling For Speech Recognition", *Language Processing For Speech Recognition*, pp. 450–506.

Abe, et al.: "Training of Lexical Models Based on DTW–Based Parameter Reestimation Algorithm", Published 1988 by IEEE, pp. 623–626.

Dialog File 2, Acc. #03157067: Snyer, et al: "Logic Programming for Speech Understanding", Published 1986 by Editions Hermes, Paris, France (Abstract Only).

Catizone, et al: "Deriving Translation Data from Bilingual Texts"; *Proceeding of the First International Acquisition Workshop*, Detroit, Mich. 1989, pp. 1–6.

US-PAT-NO: 5768603
DOCUMENT-IDENTIFIER: US 5768603 A
TITLE: Method and system for natural language translation

Abstract Text - ABTX (1):

The present invention is a system for translating text from a first source language into a second target language. The system assigns probabilities or scores to various target-language translations and then displays or makes otherwise available the highest scoring translations. The source text is first transduced into one or more intermediate structural representations. From these intermediate source structures a set of intermediate target-structure hypotheses is generated. These hypotheses are scored by two different models: a language model which assigns a probability or score to an intermediate target structure, and a translation model which assigns a probability or score to the event that an intermediate target structure is translated into an intermediate source structure. Scores from the translation model and language model are combined into a combined score for each intermediate target-structure hypothesis. Finally, a set of target-text hypotheses is produced by transducing the highest scoring target-structure hypotheses into portions of text in the target language. The system can either run in batch mode, in which case it translates source-language text into a target language without human assistance, or it can function as an aid to a human translator. When functioning as an aid to a human translator, the human may simply select from the various translation hypotheses provided by the system, or he may optionally provide hints or constraints on how to perform one or more of the stages of source transduction, hypothesis generation and target transduction.

Brief Summary Text - BSTX (6):

This impasse was overcome in the early 1970's with the introduction of statistical techniques to speech recognition. In the statistical approach, linguistic rules are extracted automatically using statistical techniques from large databases of speech and text. Different types of linguistic information are combined via the formal laws of probability theory. Today, almost all speech recognition systems are based on statistical techniques.

Brief Summary Text - BSTX (7):

Speech recognition has benefited by using statistical language models which exploit the fact that not all word sequences occur naturally with equal probability. One simple model is the trigram model of English, in which it is assumed that the probability that a word will be spoken depends only on the previous two words that have been spoken. Although trigram models are simple-minded, they have proven extremely powerful in their ability to predict words as they occur in natural language, and in their ability to improve the performance of natural-language speech recognition. In recent years more sophisticated language models based on probabilistic decision-trees, stochastic context-free grammars and automatically discovered classes of words have also been used.

Brief Summary Text - BSTX (8):

In the early days of speech recognition, acoustic models were created by linguistic experts, who expressed their knowledge of acoustic-phonetic rules in programs which analyzed an input speech signal and produced as output a sequence of phonemes. It was thought to be a simple matter to decode a word sequence from a sequence of phonemes. It turns out, however, to be a very difficult job to determine an accurate phoneme sequence from a speech signal. Although human experts certainly do exist, it has proven extremely difficult to formalize their knowledge. In the alternative statistical approach, statistical models, most predominantly hidden Markov models, capable of learning acoustic-phonetic knowledge from samples of speech are employed.

Brief Summary Text - BSTX (10):

Statistical techniques in speech recognition provide two advantages over the rule-based approach. First, they provide means for automatically extracting information from large bodies of acoustic and textual data, and second, they provide, via the formal rules of probability theory, a systematic way of combining information acquired from different sources. The problem of machine translation between natural languages is an entirely different problem than that of speech recognition. In particular, the main area of research in speech recognition, acoustic modeling, has no place in machine translation. Machine translation does face the difficult problem of coping with the complexities of natural language. It is natural to wonder whether this problem won't also yield to an attack by statistical methods, much as the problem of coping with the complexities of natural speech has been yielding to such an attack. Although the statistical models needed would be of a very different nature, the principles of acquiring rules automatically and combining them in a mathematically principled fashion might apply as well to machine translation as they have to speech recognition.

Brief Summary Text - BSTX (14):

A target-structure language model is used to estimate a first score which is proportional to the probability of occurrence of each intermediate target-structure of text associated with the target hypotheses. A target structure-to-source-structure translation model is used to estimate a second score which is proportional to the probability that the intermediate target-structure of text associated with the target hypotheses will translate into the intermediate source-structure of text. For each target hypothesis, the first and second scores are combined to produce a target hypothesis match score.

Brief Summary Text - BSTX (18):

The intermediate target structures may be expressed as an ordered sequence of morphological units, and the first score probability may be obtained by multiplying the conditional probabilities of each morphological unit within an intermediate target structure given the occurrence of previous morphological units within the intermediate target structure. In another embodiment, the conditional probability of each unit of each of the intermediate target structure may depend only on a fixed number of preceding units within the intermediate target structure.

Drawing Description Text - DRTX (66):

FIG. 64 is an example of a subset lattice;

Detailed Description Text - DETX (104):

1. A language model 101 which assigns a probability or score P(E) to any portion of English text E;

Detailed Description Text - DETX (105):

2. A translation model 102 which assigns a conditional probability or score P(F.vertline.E) to any portion of French text F given any portion of English text E; and

Detailed Description Text - DETX (106):

3. A decoder 103 which given a portion of French text F finds a number of portions of English text E, each of which has a large combined probability or score

Detailed Description Text - DETX (108):

A shortcoming of the simplified architecture depicted in FIG. 1 is that the language model 101 and the translation model 102 deal directly with unanalyzed text. The linguistic information in a portion of text and the relationships between translated portions of text is complex, involving linguistic phenomena of a global nature. However, the models 101 and 102 must be relatively simple so that their parameters can be reliably estimated from a manageable amount of training data.. In particular, they are restricted to the description of local structure.

Detailed Description Text - DETX (115):

4. an English structure language model 204 which assigns a probability or score P(E') to any intermediate

structure E';

Detailed Description Text - DETX (116):

5. an English structure to French structure translation model 205 which assigns a conditional probability or score P(F'.vertline.E') to any intermediate structure F' given any intermediate structure E'; and

Detailed Description Text - DETX (117):

6. a decoder 206 which given a French structure F' finds a number of English structures E', each of which has a large combined probability or score

Detailed Description Text - DETX (127):

In step 505 in FIG. 5, the user-aided system, portions of source text may be selected by the user. The user might, for example, select a whole document, a sentence, or a single word. The system might then show the users possible translations of the selected portion of text. For example, if t(he user selected only a single word the system might show a ranked list of possible translations of that word. The ranks being determined by statisical models that would be used to estimate the probabilities that the source word translates in various manners in the source context in which the source word appears.

Detailed Description Text - DETX (130):

FIG. 7 depicts in more detail the step 404 of the batch translation system 401 depicted in FIG. 4. The step 404 is expanded into four steps. In the first step 701 the input source text 707 is transduced to one or more intermediate source structures. In the second step 702 a set of one or more hypothesized target structures are generated. This step 702 makes use of a language model 705 which assigns probabilities or scores to target structures and a translation model 706 which assigns probabilities or scores to source structures given target structures. In the third step 703 the highest scoring hypothesis is selected. In the fourth step 704 the hypothesis selected in step 703 is transduced into text in the target language 708.

Detailed Description Text - DETX (139):

The computer platform 1014 typically includes an operating system 1003. A data storage device 1002 is also called a secondary storage and may include hard disks and/or tape drives and their equivalents. The data storage device 1002 represents non-volatile storage. The data storage 1002 may be used to store data for the language and translation models components of the translation system 1001.

Detailed Description Text - DETX (140):

Various peripheral components may be connected to the computer platform 1014, such as a terminal 1012, a microphone 1008, a keyboard 1013, a scanning device 1009, an external network 1010, and a printing device 1011. A user 503 may interact with the translation system 1001 using the terminal 1012 and the keyboard 1013, or the microphone 1008, for example. As another example, the user 503 might receive a document from the external network 1010, translate it into another language using the translation system 1001, and then send the translation out on the external network 1010.

Detailed Description Text - DETX (142):

The translation system can receive source text in a variety of known manners. The following are only a few examples of how the translation system receives source text (e.g. data), to be translated. Source text to be translated may be directly entered into the computer system via the keyboard 1013. Alternatively, the source text could be scanned in using the scanner 1009. The scanned data could be passed through a character recognizer in a known manner and then passed on to the translation system for translation. Alternatively, the user 503 could identify the location of the source text in main or secondary storage, or perhaps on removable secondary storage (such as on a floppy disk), the computer system could retrieve and then translate the text accordingly. As a final example, with the addition of a speech recognition component, it would also be possible to speak into the microphone 1008, have the speech converted into source text by the speech recognition component.

Detailed Description Text - DETX (143):

Translated target text produced by the translation application running on the computer system may be output by the system in a variety of known manners. For example, it may be displayed on the terminal 1012, stored in RAM 1005, stored data storage 1002, printed on the printer 1011, or perhaps transmitted out over the external network 1010. With the addition of a speech synthesizer it would also be possible to convert translated target text into speech in target language.

Detailed Description Text - DETX (144):

Step 403 in FIG. 4 and step 505 in FIG. 5 measure, receive or otherwise capture a portion of source text to be translated. In the context of this invention, text is used to refer to sequences of characters, formatting codes, and typographical marks. It can be provided to the system in a number of different fashions, such as on a magnetic disk, via a computer network, as the output of an optical scanner, or of a speech recognition system. In some preferred embodiments, the source text is captured a sentence at a time. Source text is parsed into sentences using a finite-state machine which examines the text for such things as upper and lower case characters and sentence terminal punctuation. Such a machine can easily be constructed by someone skilled in the art. In other embodiments, text may be parsed into units such as phrases or paragraphs which are either smaller or larger than individual sentences.

Detailed Description Text - DETX (155):

It should be understood that FIG. 11 represents only one embodiments of the source-transducer 701. Many variations are possible. For example, the transducers 1101, 1102, 1103, 1104, 1105, 1106 may be augmented and/or replaced by other transducers. Other embodiments of the source-transducer 701 may include a transducer that groups words into compound words or identifies idioms. In other embodiments, rather than a single intermediate source-structure being produced for each source sentence, a set of several intermediate source-structures together with probabilities or scores may be produced. In such embodiments the transducers depicted in FIG. 11 can be replaced bar transducers which produce at each stage intermediate structures with probabilities or scores. In addition, the intermediate source-structures produced may be different. For example, the intermediate structures may be entire parse trees, or case frames for the sentence, rather than a sequence of morphological units. In these cases, there may be more than one intermediate source-structure for each sentence with scores, or there may be only a single intermediate source-structure.

Detailed Description Text - DETX (215):

Referring again to FIG. 11, the transducer 1103 annotates words with part-of-speech labels. These labels are used by the subsequent transducers depicted the figure. In some embodiments of transducer 1103, part-of-speech labels are assigned to a word sequence using a technique based on hidden Markov models. A word sequence is assigned the most probable part-of-speech sequence according to a statistical model, the parameters of which are estimated from large annotated texts and other even larger un-annotated texts. The technique is fully explained in article by Bernard Merialdo entitled `Tagging text with a Probabilistic Model` in the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May 14-17, 1991. This article is incorporated by reference herein.

Detailed Description Text - DETX (367):

Registers: Just knowing that a regular expression matches an input attribute tuple sequence usually does not provide enough information for the construction of an appropriate output attribute tuple sequence. Data is usually also required about the attribute tuples matched by different elements of the regular expression. In ordinary LEX, to extract this type of information often requires the matched input sequence to be parsed again. To avoid this cumbersome approach, the pattern-matcher 1601 makes details about the positions in the input stream of the matched elements of the regular expression more directly available. From these positions, the identities of the attribute tuples can then be determined.

Detailed Description Text - DETX (394):

It should be understood that FIG. 19 represents only one possible embodiment of the target-transducer 704. Many variations are possible. For example, in other embodiments, rather than a single target sentence being produced for each intermediate target-structure, a set of several target sentences together with probabilities or scores may be produced. In such embodiments, the transducers depicted in FIG. 19 can be replaced by transducers which produce at each stage several target sentences with probabilities or scores. Moreover, in embodiments of the present invention in which the intermediate target-structures are more sophisticated than lexical morph sequences, the target-structure transducer is also more involved. For example, if the intermediate target-structure consists of a parse tree of a sentence or case frames for a sentence, then the target-structure transducer converts these to the target language.

Detailed Description Text - DETX (441):

To perform such conversions, the transducer 1909 uses a target-language model to assign a probability or score to each of the different possible sentences corresponding to an input sentence with a morphological ambiguity. The sentence with the highest probability or score is selected. In the above example, the sentence how quickly can you run? is selected because it has a higher target-language model probability or score than how quick can you run? In some embodiments of the transducer 1909 the target-language model is a trigram model similar to the target-structure language model 706. Such a transducer can be constructed by a person skilled in the art. The last transducer 1910 assigns a case to the words of a target sentence based on the casing rules for English. Principally this involves capitalizing the words at the beginning of sentences. Such a transducer can easily be constructed by a person skilled in the art.

Detailed Description Text - DETX (447):

The inventions described in this specification employ probabilistic models of the target language in a number of places. These include the target structure language model 705, and the class language model used by the decoder 404. As depicted in FIG. 20, the role of a language model is to compute an a priori probability or score of a target structure.

Detailed Description Text - DETX (453):

Since it is expensive to evaluate a language model in the context of a complete system, it is useful to have an intrinsic measure of the quality of a language model. One such measure is the probability that the model assigns to the large sample of target structures. One judges as better the language model which yields the greater probability. When the target structure is a sequence of words or morphs, this measure can be adjusted so that it takes account of the length of the structures. This leads to the notion of the perplexity of a language model with respect to a sample of text S: ##EQU1## where .vertline.S.vertline. is the number of morphs of S. Roughly speaking, the perplexity is the average number of morphs which the model cannot distinguish between, in predicting a morph of S. The language model with the smaller perplexity will be the one which assigns the larger probability to S.

Detailed Description Text - DETX (456):

n-gram language models will now be described. For these models, the target structure consists of a sequence of morphs. Suppose m.sub.1 m.sub.2 m.sub.3 . . . be a sequence of k morphs m.sub.i. For 1.ltoreq.i.ltoreq.j.ltoreq.k, let m.sub.i.sup.j denote the subsequence m.sub.i.sup.j .ident.m.sub.i m.sub.i+1 . . . m.sub.j. For any sequence, the probability of a m.sub.1.sup.k .ident.is equal to the product of the conditional probabilities of each morph .sub.i given the previous morphs m.sub.1.sup.i-1 :

Detailed Description Text - DETX (458):

For an n-gram model, the conditional probability of a morph in a sequence is assumed to depend on its history only through the previous n-1 morphs:

Detailed Description Text - DETX (459):

For a vocabulary of size V, a 1-gram model is determined by V-1 independent numbers, one probability Pr(m) for each morph m in the vocabulary, minus one for the constraint that all of the probabilities add up to 1. A 2-gram model is determined by V.sup.2 -1 independent numbers, V(V-1) conditional probabilities of the form Pr (m.sub.2 .vertline.m.sub.1) and V-1 of the form Pr (m). In general, an n-gram model is determined by V.sup.n -1 independent numbers, V.sup.n-1 (V-1) conditional probabilities of the form Pr(m.sub.n .vertline.m.sub.1.sup.n-1), called the order-n conditional probabilities, plus V.sup.n-1 -1 numbers which determine an (n-1)-gram model.

Detailed Description Text - DETX (460):

The order-n conditional probabilities of an n-gram model form the transition matrix of an associated Markov model. The states of this Markov model are sequences of n-1 morphs, and the probability of a transition from the state m.sub. m.sub.2 . . . m.sub.n-1 . . . to the state m.sub.2 m.sub.3 . . . m.sub.n is Pr(m.sub.n .vertline.m.sub.1 m.sub.2 . . . m.sub.n-1). An n-gram language model is called consistent if, for each string m.sub.1.sup.n-1, the probability that the model assigns to m.sub.1.sup.n-1 is the steady state probability for the state m.sub.1.sup.-1 of the associated Markov model.

Detailed Description Text - DETX (462):

The simplest form of an n-gram model is obtained by assuming that all the independent conditional probabilities are independent parameters. For such a model, values for the parameters can be determined from a large sample of training text by sequential maximum likelihood training. The order n-probabilities are given by ##EQU2## where f(m.sub.n .vertline.m.sub.1.sup.n-) is the number of times the string of morphs m.sub.1 m.sup.i appears in the training text.

Detailed Description Text - DETX (464):

Unfortunately, many of the parameters of a simple n-gram model will not be reliably estimated by this method. The problem is illustrated in Table 3, which shows the number of 1-, 2-, and 3-grams appearing with various frequencies in a sample of 365,893,263 words of English text from a variety of sources. The vocabulary consists of the 260,740 different words plus a special unknown word into which all other words are mapped. Of the 6.799.times.10.sup.10 2 -grams that might have occurred in the data, only 14,494,217 actually did occur and of these, 8,045,024 occurred only once each. Similarly, of the 1.773.times.10.sup.16 3-grams that might have occurred, only 75,349,888 actually did occur and of these, 53,737,350 occurred only once each. These data and Turing's formula imply that 14.7 percent of the 3-grams and for 2.2 percent of the 2-grams in a new sample of English text will not appear in, the original

Detailed Description Text - DETX (466):

sample. Thus, although any 3-gram that does not appear in the original sample is rare, there are so many of them that their aggregate probability is substantial.

Detailed Description Text - DETX (469):

A solution to this difficulty is provided by interpolated estimation, which is described in detail in the paper "Interpolated estimation of Markov source parameters from sparse data", by F. Jelinek and R. Mercer and appearing in Proceeding of the Workshop on Pattern Recognition in Practice, published by North-Holland, Amsterdam, The Netherlands, in May 1980. Interpolated estimation combines several models into a smoothed model which uses the probabilities of the more accurate models where they are reliable and, where they are unreliable, falls back on the more reliable probabilities of less accurate models. If Pr.sup.(j) (m.sub.i .vertline.m.sub.1.sup.i-1)is the jth language model, the smoothed model, Pr (m.sub.i .vertline.m.sub.1.sup.i-1), is given by ##EQU3## The value of the .lambda..sub.j (m.sub.1.sup.i-1) are determined using the EM method, so as to maximize the probability of some additional sample of training text called held-out data. When interpolated estimation is used to combine simple 1-, 2-, and 3-gram models, the .lambda.'s can be chosen to depend on m.sub.1.sup.i-1 only through the count of m.sub.i-2 m.sub.i-1. Where this count is high, the simple 3-gram model will be reliable, and, where this count is low, the simple 3-gram model will be unreliable.

Detailed Description Text - DETX (470):

The inventors constructed an interpolated 3-gram model in which the .lambda.'s were divided into 1782 different sets according to the 2-gram counts, and determined from a held-out sample of 4,630,934 million words. The power of the model was tested using the 1,014,312 word Brown corpus. This well known corpus, which contains a wide variety of English text, is described in the book Computational Analysis of Present-Day American English, by H. Kucera and W. Francis, published by Brown University Press, Providence, R.I., 1967. The Brown corpus was not included in either the training or held-out data used to construct the model. The perplexity of the interpolated model with respect to the Brown corpus was 244.

Detailed Description Text - DETX (472):

Clearly, some words are similar to other words in their meaning and syntactic function. For example, the probability distribution of words in the vicinity of Thursday is very much like that for words in the vicinity of Friday. Of course, they will not be identical: people rarely say Thank God it's Thursday! or worry about Thursday the 13.sup.th.

Detailed Description Text - DETX (475):

In a simple n-gram class model, the C.sup.n .times.1+V-C independent probabilities are treated as independent parameters. For such a model, values for the parameters can be determined by sequential maximum likelihood training. The order n probabilities are given by ##EQU4## where f(c.sub.1.sup.i) is the number of times that the sequence of classes c.sub.1.sup.i appears in the training text. (More precisely, f(c.sub.1.sup.i) is the number of distinct occurrences in the training text of a

Detailed Description Text - DETX (480):

A general scheme for clustering a vocabulary into classes is depicted schematically in FIG. 31. It takes as input a desired number of classes C 3101, a vocabulary 3102 of size V, and a model 3103 for a probability distribution P(w.sub.1,w.sub.2) over bigrams from the vocabulary. It produces as output a partition 3104 of the vocabulary into C classes. In one application, the model 3103 can be a 2-gram language model as described in Section 6, in which case P(w.sub.1,w.sub.2) would be proportional to the number of times that the bigram w.sub.1 w.sub.2 appears in a large corpus of training text.

Detailed Description Text - DETX (481):

Let the score .psi.(C) of a partition C be the average mutual information between the classes of C with respect to the probability distribution P(w.sub.1,w.sub.2): ##EQU5## In this sum, c.sub.1 and c.sub.2 each run over the classes of the partition C, and ##EQU6## The scheme of FIG. 31 chooses a partition C for which the score average mutual information .psi.(C) is large.

Detailed Description Text - DETX (492):

The implementation can improved further by keeping track of those pairs l,m, for which p.sub.k (l,m) is different from zero. For example, suppose that P is given by a simple bigram model trained on the data described in Table 3 of Section 6. In this case, of the 6.799.times.10.sup.10 possible word 2-grams w.sub.1,w.sub.2, only 14,494,217 have non-zero probability. Thus, in this case, the sums required in equation 18 have, on average, only about 56 non-zero terms instead of 260,741 as might be expected from the size of the vocabulary.

Detailed Description Text - DETX (497):

7.5 Examples The methods described above were used divide the 260,741-word vocabulary of Table 3, Section 6, into 1000 classes. Table 4 shows some of the classes that are particularly interesting, and Table 5 shows classes that were selected at random. Each of the lines in the tables contains members of a different class. The average class has 260 words. The table shows only those words that occur at least ten times, and only the ten most frequent words of any class. (The other two months would appear with the class of months if this limit had been extended to twelve). The degree to which the classes capture both syntactic and semantic aspects of English is quite surprising given that they were constructed from nothing more than counts of bigrams. The class [that tha theat] is interesting because although tha and theat are English words, the method has discovered that in the training data cach of them is most often a mistyped that.

Detailed Description Text - DETX (544):

As illustrated in FIG. 21, a target structure to source structure translation model P.sub..theta. 706 with parameters .theta. is a method for calculating a conditional probability, or likelihood, P.sub..theta. ((f.vertline.e), for any source structure f given any target structure e. Examples of such structures include, but are not limited to, sequences of words, sequences of linguistic morphs, parse trees, and case frames. The probabilities satisfy: ##EQU11## where the sum ranges over all structures f, and failure is a special symbol. P.sub..theta. (f.vertline.e) can be interpretted as the probability that a translator will produce f when given e, and P.sub..theta. (failure.vertline.e) can be interpreted as the probability that he will produce no translation when given e. A model is called deficient if P.sub..theta. (failure.vertline.e) is greater than zero for some e.

Detailed Description Text - DETX (547):

One training methodology is maximum likelihood training, in which the parameter values are chosen so as to maximize the probability that the model assigns to a training sample consisting of a large number S of translations (f.sup.(s),e.sup.(s)), s=1,2, . . . ,S. This is equivalent to maximizing the log likelihood objective function ##EQU12## Here C(f,e) is the empirical distribution of the sample, so that C(f,e) is 1/S times the number of times (Usually 0 or 1) that the pair (f,e) occurs in the sample.

Detailed Description Text - DETX (551):

In some embodiments, illustrated in FIG. 24, a translation model 706 computes the probability of a source structure given a target structure as the sum of the probabilities of all alignments between these structures: ##EQU13## In other embodiments, a translation model 706 can compute the probability of a source structure given a target structure as the maximum of the probabilities of all alignments between these structures: ##EQU14## As depicted in FIG. 25, the probability P.sub..theta. (f.vertline.e) of a single alignment is computed by a detailed translation model 2101. The detailed translation model 2101 employs a table 2501 of values for the parameters .theta..

Detailed Description Text - DETX (558):

The probability of an alignment and source structure given a target structure is obtained by combining the probabilities computed by each of these sub-models. Corresponding to these sub-models, the table of parameter values 2501 comprises:

Detailed Description Text - DETX (559):

1a fertility probabilities n(.o slashed..vertline.e), where .o slashed. is any non-negative integer and e is any target morph;

Detailed Description Text - DETX (560):

b. null fertility probabilities n.sub.0 (.o slashed..vertline.m'), where .o slashed. is any non-negative integer and m' is any positive integer;

Detailed Description Text - DETX (561):

2a lexical probabilities t(f.vertline.e), where f is any source morph, and e is any target morph;

Detailed Descripti n Text - DETX (562):

b. lexical probabilities t(f.vertline.*null*), where f is any source morph, and *null* is a special symbol;

Detailed Description Text - DETX (563):

3. distortion probabilities a(j.vertline.i,m), where m is any positive integer, i is any positive integer, and j is any positive integer between 1 and m.

Detailed Description Text - DETX (564):

This embodiment of the detailed translation model 2101 computes the probability P.sub..theta. (f,a.vertline.e) of an alignment a and a source structure f given a target structure e as follows. If any source entry is connected to more than one target entry, then the probability is zero. Otherwise, the probability is computed by the formula

Detailed Description Text - DETX (576):

The components of the probability P.sub..theta. (f,a.vertline.e) are ##EQU16## Many other embodiments of the detailed translation 2101 are possible. Five different embodiments will be described in Section 9 and Section 10.

Detailed Description Text - DETX (588):

Two hidden alignment models P.sub..theta. and P.sub..theta. of the form depicted in FIG. 24 can be compared using the relative objective function .sup.1 .sup.1 The relative objective function R(P.sub..theta.,P.sub..theta.) is similar in form to the relative entropy ##EQU17## between probability distributions p and q. However, whereas the relative entropy is never negative, R can take any value. The inequality (35) for R is the analog of the inequality D.gtoreq.0 for D. ##EQU18## where P.sub..theta. (a.vertline.f,e)=P.sub..theta. (a,f.vertline.e)/ P.sub..theta. (f.vertline.e). Note that R(P.sub..theta., P.sub..theta.)=0. R is related to .psi. by Jensen's inequality

Detailed Description Text - DETX (604):

To apply these procedures, it is necessary to solve the maximization problems of Steps 2802 and 2901. For the models described below, this cain be done explicitly. To see the basic form of the solution, suppose P.sub..theta. is a simple model given by ##EQU20## where the .theta.(.omega.),.omega..epsilon..OMEGA., are real valued parameters satisfying the constraints ##EQU21## and for each .omega. and (a,f,e), c(.omega.; a,f,e) is a non-negative integer..sup.2 It is natural to interpret .theta.(.omega.) as the probability of the event .omega. and c (.omega.; a,f,e) as the number of times that this event occurs in (a,f,e). Note that ##EQU22## .sup.2 More generally, we can allow c(.omega.;a,f,e) to be a non-negative real number.

Detailed Description Text - DETX (606):

These formulae can easily be generalized to models of the form (38) for which the single equality constraint in Equation (39) is replaced by multiple constraints ##EQU25## where the subsets .OMEGA..sub..mu.,.mu.=1,2, . . . form a partition of .OMEGA.. Only Equation (42) needs to be modified to include a different .lambda..sub..mu. for each .mu.: if .psi..epsilon..OMEGA..sub..mu., then ##EQU26## 8.3.5 Approximate and Viterbi Parameter Estimation

Detailed Description Text - DETX (608):

For simple models, such as Model 1 and Model 2 described below, it is possible to calculate these counts exactly by including the contribution of each possible alignments. For more sophisticated models, such as Model 3 Model 4, and Model 5 described below, the sum over alignments in Equation 44 is too costly to compute exactly. Rather, only the contributions from a subset of alignments can be practically included. If the alignments in this subset account for most of the probability of a translation, then this truncated sum can still be a good approximation.

Detailed Description Text - DETX (610):

In calculating the counts using the update formulae 44, approximate the sum by including only the contributions from some subset of alignments of high probability.

Detailed Description T xt - DETX (611):

If only the contribution from the single most probable alignment is included, the resulting procedure is called Viterbi Parameter Estimation. The most probable alignment between a target structure and a source structure is called the Viterbi alignment. The convergence of Viterbi Estimation is easily demonstrated. At each iteration, the parameter values are re-estimated so as to make the current set of Viterbi alignments as probable as possible; when these parameters are used to compute a new set of Viterbi alignments, either the old set is recovered or a set which is yet more probable is found. Since the probability can never be greater than one, this process surely converge. In fact, it converge in a finite, though very large, number of steps because there are only a finite mummer of possible alignments for any particular translation.

Detailed Description Text - DETX (615):

Model 1 is very simple but it is useful because its likelihood function is concave and consequently has a global maximum which can be found by the EM procedure. Model 2 is a slight generalization of Model 1. For both Model 1 and Model 2, the sum over alignments for the objective function and the relative objective function can be computed very efficiently. This significantly reduces the computational complexity of training. Model 3 is more complicated and is designed to more accurately model the relation between a morph of e and the set of morphs in f to which it is connected. Model 4 is a more sophisticated step in this direction. Both Model 3 and Model 4 are deficient. Model 5 is a generalization of Model 4 in this deficiency is removed at the expense of more increased complexity. For Models 3,4, and 5 the exact sum over alignments can not be computed efficiently. Instead, this sum can be approximated by restricting it to alignments of high probability.

Detailed Description Text - DETX (618):

In this section embodiments of the statistical translation model that assigns a conditional probability to the event that a sequence of lexical units in the source language is a translation of a sequence of lexical units in the target language will be described. Methods for estimating the parameters of these embodiments will be explained. For concreteness the discussion will be phrased in terms of a source language of French and a target language of English. The sequences of lexical units will be restricted to sequences of words.

Detailed Description Text - DETX (621):

Random variables will be denoted by upper case letters, and the values of such variables will be denoted by the corresponding lower case letters. For random variables X and Y, the probability Pr(Y=y.vertline.X=x) will be denoted by the shorthand P(y.vertline.x). Strings or vectors will be denoted by bold face letters, and their entries will be denoted by the corresponding non-bold letters.

Detailed Description Text - DETX (628):

The alignment in FIG. 33 has seven connections, (the, Le), (program, programme), and so on. In the description that follows, such an alignment will be denoted as (Le programme a ete mis en application.vertline.And the(1) program(2)has(3)been(4) implemented(5,6,7)). The list of numbers following an English word shows the positions in the French string of the words with which it is aligned. Because And is not aligned with any French words here, there is no list of numbers after it. Every alignment is considered to be correct with some probability. Thus (Le programme a ete mis en application.vertline.And (1,2,3,4,5,6,7) the programme has been implemented) is perfectly acceptable. Of course, this is much less probable than the alignment show in FIG. 33.

Detailed Description Text - DETX (630):

The set of English words connected to a French word will be called the notion that generates it. An alignment resolves the English string into a set of possibly overlapping notions that is called a notional scheme. The previous example contains the three notions The, poor, and don't have any money. Formally, a notion is a subset of the positions in the English string together with the words occupying those positions. To avoid confusion in describing such cases, a subscript will be affixed to each word showing its position. The alignment in FIG. 34 includes the notions the.sub.4 and of.sub.6 the .sub.7, but not the notions of.sub.6 the .sub.4 or the.sub.7. In (J'applaudis ala decision.vertline.I(1) applaud(2) the(4) decision(5)), a is generated by the empty notion. Although the empty notion has no position and so never requires a clarifying subscript, it will be placed at the beginning of the English string, in position zero, and denoted by e.sub.0. At times, therefore, the previous alignment will be denoted as (J'aplaudis a la decision.vertline.e.sub.0 (3) I(1) applaud(2) the(4) decision(5)).

Detailed Description Text - DETX (631):

The set of all alignments of (f.vertline.e) will he written A(e,f). Since e has length I and f has length m, there are Im different connections that can be drawn between them. Since an alignment is determined by the connections that it contains, and since a subset of the possible connections can be chosen in 2.sup.Im ways, there are 2.sup.Im alignments in A(e,f).

Detailed Description Text - DETX (633):

The probability of a French string f and an alignment a given an English string e can be written ##EQU27## Here, m is the length of f and a.sub.1.sup.m is determined by a. 9.4 Model 1

Detailed Description Text - DETX (634):

The conditional probabilities on the right-hand side of Equation (47) cannot all be taken as independent parameters because there are too many of them. In Model 1, the probabilities .sub.P (m.vertline.e) are taken to be independent of e and m; that .sub.P (a.sub.j .vertline.a.sub.1.sup.j-1,f.sub.1.sup.j-,m,e), depends only on I, the length of the English string, and therefore must be (I+1).sup.-1 ; and that P(f.sub.j .vertline.a.sub.1.sup.j, m,e), depends only on f.sub.j and e.sub.2.sbsb.j. The parameters, then, are .epsilon..ident..sub.P (m.vertline.e), and t (f.sub.j .vertline.e.sub.a.sbsb.j).ident.P(f.sub.j .vertline.a.sub.1hu j-1,m, e), which will be called the translation probability of f.sub.j given e.sub.a.sbsb.j.

Detailed Description Text - DETX (635):

The parameter .epsilon. is fixed at some small number. The distribution of M is unnormalized but this is a minor technical issue of no significance to the computations. In particular, M, can be restricted to some finite range. As long as this range encompasses everything that actually occurs in training data no problems arise.

Detailed Description Text - DETX (636):

A method for estimating the translation probabilities for Model 1 will now be described.

Detailed Description Text - DETX (638):

The first goal of the training process is to find the values of the translation probabilities that maximize p (f.vertline.e) subject to the constraints that for each e, ##EQU30## An iterative method for doing so will be described.

Detailed Description Text - DETX (639):

The method is motivated by the following consideration. Following standard practice for constrained maximization, a necessary relationship for the parameter values at a, local maximum can be found by introducing Lagrange multipliers, .lambda.e, and seeking an unconstrained maximum of the auxiliary function ##EQU31## At a local maximum, all of the partial derivatives of h with respect to the components of t and .lambda. are zero. That the partial derivatives with respect to the components of .lambda. be zero is simply a restatement of the constraints on the translation probabilities. The partial derivative of h with respect to t(f.vertline.e) is ##EQU32## where .delta. is the Kronecker delta function, equal to one when both of its arguments are the same and equal to zero otherwise. This will be zero provided that ##EQU33## Superficially Equation (53) looks like a, solution to the maximization problem, but it is not because the translation probabilities appear on both sides of the equal sign. Nonetheless, it suggests an iterative procedure for finding a solution:

Detailed Description Text - DETX (640):

1. Begin with initial guess for the translation probabilities;

Detailed Description Text - DETX (644):

(Here and elsewhere, the Lagrange multipliers simply serve as a reminder that the translation probabilities must be normalized so that they satisfy Equation (50).) This process, when applied repeatedly is called the EM process. That it converges to a stationary point of h in situations like this, as demonstrated in the previous section, was first shown by L. E. Baum in an article entitled, An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process, appearing in the journal Inequalities, Vol.3, in 1972.

Detailed Description Text - DETX (646):

In practice, the training data consists of a set of translations, (f.sup.(1) .vertline.e.sup.(1)), (f.sup.(2) .vertline.e.sup.(2)), . . . , (f.sup.(S) .vertline.e.sup.(S)), so this equation becomes ##EQU36##

Detailed Description Text - DETX (647):

Here, .lambda..sub.e serves only as a reminder that the translation probabilities must be normalized.

Detailed Description Text - DETX (648):

Usually, it is not feasible to evaluate the expectation in Equation (55) exactly. Even if multiword notions are excluded, there are still (l+1).sup.m alignments possible for (f.vertline.e). Model 1, however, is special because by recasting Equation (49), it is possible to obtain an expression that can be evaluated efficiently. The right-hand side of Equation (49) is a sum of terms each of which is a monomial in the translation probabilities. Each monomial contains m translation probabilities, one for each of the words in f. Different monomials correspond to different ways of connecting words in f to notions in e with every way appearing exactly once. By direct evaluation, then ##EQU37## Therefore, the sums in Equation (49) can be interchanted with the product to obtain ##EQU38## Using this expression, it follows that ##EQU39## Thus, the number of operations necessary to calculate a count is proportional to lm rather than to (l+1).sup.m as Equation (55) might suggest.

Detailed Description Text - DETX (649):

The details of the initial guesses for t(f.vertline.e) are unimportant because P(f.vertline.e) has a unique local maximum for Model 1, as is shown in Section 10. In practice, the initial probabilities t(f.vertline.e) are chosen to be equal, but any other choice that avoids zeros would lead to the same final solution.

Detailed Description Text - DETX (651):

Model 1, takes no cognizance of where words appear in either string. The first word in the French string is just as likely to be connected to a word at the end of the English string as to one at the beginning. In contrast, Model 2 makes the same assumptions as in Model 1 except that it assumes that P(a.sub.j .vertline.a.sub.1.sup.j-1,f.sub.1.sup.j-1, m, e) depends on j, a.sub.j, and m, as well as on l. This is done using a set of alignment probabilities.

Detailed Description Text - DETX (653):

and, for a set of translations, ##EQU44## Notice that if f.sup.(s) does not have length m or if e.sup.(s) does not have length l, then the corresponding count is zero. As with the .lambda.'s in earlier equations, the .mu.'s here serve to normalize the alignment probabilities.

Detailed Description Text - DETX (654):

Model 2 shares with Model 1 the important property that the sums in Equations (55) and (65) can be obtained efficiently. Equation (63) can be rewritten ##EQU45## Using this form for P(f.vertline.e), it follows that ##EQU46## Equation (69) has a double sum rather than the product of two single sums, as in Equation (60), because, in Equation (69), i and j are tied together through the alignment probabilities.

Detailed Description Text - DETX (656):

Taking as initial estimates of the parameters for Model 2 the parameter values that result from training Model 1 is equivalent to computing the probabilities of all alignments using Model 1, but then collecting the counts appropriate to Model 2. The idea of computing the probabilities of the alignments using one model, but collecting the counts in a way appropriate to a second model is very general and can always be used to transfer a set of parameters from one model to another.

Detailed Description Text - DETX (658):

Models 1 and 2 make various approximations to the conditional probabilities that appear in Equation (47). Although Equation (47) is an exact statement, it is only one of many ways in which the joint likelihood of f and a can be written as a product of conditional probabilities. Each such product corresponds in a natural way to a generative process for developing f and a from e. In the process corresponding to Equation (47), a length for f is chosen first. Next, a position in e is selected and connected to the first position in f. Next, the identity of f.sub.1 is selected. Next, another position in e is selected and this is connected to the second word in f, and so on.

Detailed Description Text - DETX (666):

Model 3 is based on Equation (71). It assumes that, for i between 1 and l, P(.o slashed..sub.i .vertline..o slashed..sub.1.sup.i-1,e) depends only on is and .o slashed..sub.i and that, for all i, P (.tau..sub.ik .vertline..tau..sub.i1.sup.k-1, .tau..sub.0.sup.i-1,.o slashed..sub.0.sup.l,e) depends only on .tau.v.sub.ik and e.sub.i ; and that, for i between 1 and l, P(.pi..sub.ik .vertline..pi..sub.i1.sup.k-1, .pi..sub.1.sup.i-1, .tau..sub.0.sup.l, .o slashed..sub.0.sup.l, e) depends only on .pi..sub.ik, i, m, and l. The parameters of Model 3 are thus a, set of fertility probabilities, n(.o slashed..vertline.e.sub.i).tbd.P(.o slashed..vertline..o slashed..sub.1.sup.i-1, e); a set of translation probabilities, t(f.vertline.e.sub.i).tbd.Pr(T.sub.ik =f.vertline..tau..sub.i1.sup.k-1, .tau..sub.0.sup.i-1, .o slashed..sub.0.sup.l, e); and a set of distortion probabilities, d (j.vertline.i, m, l).tbd.Pr(II.sub.ik =j.vertline..pi..sub.i1.sup.k-1, .pi..sub.1.sup.i-1, .tau..sub.0.sup.l,.o slashed..sub.0.sup.l, e).

Detailed Description Text - DETX (667):

The distortion and fertility probabilities for e.sub.0 are handled differently. The empty notion conventionally occupies position 0, but actiually has no position. Its purpose is to account for those words in the French string that cannot readily be accounted for by other notions in the English string. Because these words are usually spread uniformly throughout the French string, and because they are placed only after all of the other words in the sentence have been placed, the probability Pr(II.sub.0k+1 =j.vertline..pi..sub.01.sup.k, .pi..sub.1.sup.l, .tau..sub.0.sup.l, .o slashed..sub.0.sup.l, e) is set to 0 unless position j is vacant in which case it is set (.o slashed..sub.0 -k).sup.-1. Therefore, the contribution of the distortion probabilities for all of the words in .tau..sub.0 is 1/.o slashed..sub.0 !.

Detailed Description Text - DETX (668):

The value of .o slashed..sub.0 depends on the length of the French sentence since longer sentences have more of these extraneous words. In fact Model 3 assumes that ##EQU50## for some pair of auxiliary parameters p.sub.0 and p.sub.1. The expression on the left-hand side of this equation depends on .o slashed..sub.1.sup.l only through the sum .o slashed..sub.1 + . . . +.o slashed..sub.l and defines a probability distribution over .o slashed..sub.0 whenever p.sub.0 and p.sub.1 are nonnegative and sum to 1. The probability P(.o slashed..sub.0 .vertline..o slashed..sub.1.sup.l, e) can be interpretted as follows. Each of the words from .tau..sub.1.sup.l is imagined to require an extraneous word with probability p.sub.1 ; this word is required to be connected to the empty notion. The probability that exactly .o slashed..sub.0 of the words from .tau..sub.1.sup.l will require an extraneous word is just the expression given in Equation (73).

Detailed Description Text - DETX (669):

As in Models 1 and 2, an alignment of (f.vertline.e) in Model 3 is determined by specifying a.sub.j for each position in the French string. The fertilities, .o slashed..sub.0 through .o slashed..sub.l, are functions of the a.sub.j 's. Thus, .o slashed..sub.i is equal to the number of j's for which a.sub.j equals i. Therefore, ##EQU51## with .SIGMA..sub.f t(f.vertline.e)=1, .SIGMA..sub.j d(j.vertline.i, m, l)=1, .SIGMA..sub..o slashed. n (.o slashed..vertline.e)=1, and p.sub.0 +p.sub.1 =1. According to the assumptions of Model 3, each of the pairs (.tau.,

.pi.) in (f, a) makes an identical contribution to the sum in Equation(72). The factorials in Equation (74) come from carrying out this sum explicitly. There is no factorial for the empty notion because it is exactly cancelled by the contribution from the distortion probabilities.

Detailed Description Text - DETX (672):

To define the subset, S, of the elements of A(f.vertline.e) over which the sums are evaluated a little more notation is required. Two alignments, a and a' will be said to differ by a move if there is exactly one value of j for which a.sub.j .noteq.a.sub.j '. Alignments will be said to differ by a swap if a.sub.j =a.sub.j ' except at two values, j.sub.1 and j.sub.2, for which a.sub.j1 =a.sub.j2 ' and a.sub.j2 =a.sub.j1 '. The two alignments will be said to be neighbors if they are identical or differ by a move or by a swap. The set of all neighbors of a will be denoted by N (a).

Detailed Description Text - DETX (673):

Let b(a) be that neighbor of a for which the likelihood is greatest. Suppose that ij is pegged for a. Among the neighbors of a for which ij is also pegged, let b.sub.i.rarw.j (a) be that for which the likelihood is greatest. The sequence of alignments a, b(a), b.sup.2 (a).tbd.b(b(a)), . . . , converges in a finite number of steps to an alignment that will be denoted as b.sup..infin. (a). Similarly, if ij is pegged for a, the sequence of alignments a, b.sub.i.rarw.j (a), b.sub.i.rarw.j.sup.2 (a), . . . , converges in a finite number of steps to an alignment that will be denoted as b.sub.i.rarw.j.sup.28 (a). The simple form of the distortion probabilities in Model 3 makes it easy to find b(a) and b.sub.i.rarw.j (a). If a' is a neighbor of a obtained from it by the move of j from i to i', and if neither i nor i' is 0, then ##EQU55## Notice that .o slashed..sub.i' is the fertility of the word in position i' for alignment a. The fertility of this word in alignment a' is .o slashed..sub.i' +1. Similar equations can be easily derived when either i or i' is zero, or when a and a' differ by a swap.

Detailed Description Text - DETX (676):

In one iteration of the EM process for Model 3, the counts in Equations (76) through (80), are computed by summing only over elements of S. These counts are then used in Equations (81) through (84) to obtain a new set of parameters. If the error made by including only some of the elements of A(e, f) is not too great, this iteration will lead to values of the parameters for which the likelihood of the training data is at least as large as for the first set of parameters.

Detailed Description Text - DETX (677):

The initial estimates of the parameters of Model 2 are adapted from the final iteration of the EM process for Model 2. That is, the counts in Equations (76) through (80) are computed using Model 2 to evaluate P(a.vertline.e, f). The simple form of Model 2 again makes exact calculation feasible. The Equations (69) and (70) are readily adapted to compute counts for the translation and distortion probabilities; efficient calculation of the fertility counts is more involved. A discussion of how this is done is given in Section 10.

Detailed Description Text - DETX (679):

A problem with the parameterization of the distortion probabilities in Model 3 is this: whereas the sum over all pairs .tau., .pi. of the expression on the right-hand side of Equation (71) is unity, if Pr(II.sub.ik =j.vertline..pi..sub.i1.sup.k-1, .pi..sub.1.sup.i-1, .tau..sub.0.sup.I, .o slashed..sub.0.sup.I, e) depends only on j, i, m, and I for i >0.

Detailed Description Text - DETX (680):

Because the distortion probabilities for assigning positions to later words do not depend on the positions assigned to earlier words, Model 3 wastes some of its probability on what will be called generalized strings, i.e., strings that have some positions with several words and others with none. When a model has this property of not concentrating all of its probability on events of interest, it will be said to be deficient. Deficiency is the price for the simplicity that permits Equation (85).

Detailed Description T xt - DETX (681):

Deficiency poses no serious problem here. Although Models 1 and 2 are not technically deficient, they are surely spiritually deficient. Each assigns the same probability to the alignments (Je n'ai pas de stylo .vertline. I(1) do not(2,4) have(3) a(5) pen(6)) and (Je pas ai ne de stylo .vertline. I(1) do not(2,4) have(3) a(5) pen(6)), and, therefore, essentially the same probability to the translations (Je n'ai pas de stylo .vertline. I do not have a pen) and (Je pas ai ne de stylo .vertline. I do not have a pen). In each case, not produces two words, ne and pas, and in each case, one of these words ends up in the second position of the French string and the other in the fourth position. The first translation should be much more probable than the second, but this defect is of little concern because while the system may be required to translate the first string someday, it will almost surely not be required to translate the second. The translation models are not used to predict French given English but rather as a component of a system designed to predict English given French. They need only be accurate to within a constant factor over well-formed strings of French words.

Detailed Description Text - DETX (683):

Often the words in an English string constitute phrases that are translated as units into French. Sometimes, a translated phrase may appear at a spot in the French string different from that at which the corresponding English phrase appears in the English string. The distortion probabilities of Model 3 do not account well for this tendency of phrases to move around as units. Movement of a long phrase will be much less likely than movement of a short phrase because each word must be moved independently. In Model 4, the treatment of Pr(II.sub.ik =j.vertline..pi..sub.i1.sup.k-1, .pi..sub.1.sup.i-1, .tau..sub.0.sup.I, .o slashed..sub.0.sup.I, e) is modified so as to alleviate this problem. Words that are connected to the empty notion do not usually form phrases and so Model 4 continues to assume that these words are spread uniformly throughout the French string.

Detailed Description Text - DETX (685):

In Model 4, the probabilities d(j.vertline.i, m, l) are replaced by two sets of parameters: one for placing the head of each notion, and one for placing any remaining words. For [i]>0, Model 4 requires that the head for notion i be .tau..sub.[i]1. It assumes that

Detailed Description Text - DETX (686):

Here, A and 1 are functions of the English and French word that take on a small number of different values as their arguments range over their respective vocabularies. In the Section entitled Classes, a process is described for dividing a vocabulary into classes so as to preserve mutual information between adjacent classes in running text. The classes A and B are constructed as functions with fifty distinct values by dividing the English and French vocabularies each into fifty classes according to this method. The probability is assumed to depend on the previous notion and on the identity of the French word being placed so as to account for such facts as the appearance of adjectives before nouns in English but after them in French. The displacement for the head of notion i is denoted by j-.circle-w/dot..sub.i-1. It may be either positive or negative. The probability d.sub.1 (-1.vertline.A(e), B(f)) is expected to be larger than d.sub.1 (+1.vertline.A(e), B(f)) when e is an adjective and f is a noun. Indeed, this is borne out in the trained distortion probabilities for Model 4, where d.sub.1 (-1.vertline.A (government's), B(developpement)) is 0.9269, while d.sub.1 (+1A(government's), B(developpement)) is 0.0022.

Detailed Description Text - DETX (690):

The count and reestimation formulae for Model 4 are similar to those for the previous Models and will not be given explicitly here. The general formulae in Section 10 are helpful in deriving these formulae. Once again, the several counts for a translation are expectations of various quantities over the possible alignments with the probability of each alignment computed from an earlier estimate of the parameters. As with Model 3, these expectations are computed by sampling some small set, S, of alignments. As described above, the simple form that for the distortion probabilities in Model 3, makes it possible to find b.sup..infin. (a) rapidly for any a. The analogue of Equation (85) for Model 4 is complicated by the fact that when a French word is moved from notion to notion, the centers of two notions change, and the contribution of several words is affected. It is nonetheless possible to evaluate the adjusted likelihood incrementally, although it is substantially more time consuming.

Detail d Description Text - DETX (697):

Again, the final factor enforces the constraint that .tau..sub.[i]k land in a vacant position, and, again, it is assumed that the probability depends on f.sub.j only through its class.

Detailed Description Text - DETX (701):

A large collection of training data is used to estimate the parameters of the five models described above. In one embodiment of these models, training data is obtained using the method described in detail in the paper, Aligning Sentences in Parallel Corporation, by P. F. Brown, J. C. Lai, and R. L. Mercer, appearing in the Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Jun. 1991. This paper is incorporated by reference herein. This method is applied to a large number of translations from several years of the proceedings of the Canadian parliament. From these translations, a trailing data set is chosen comprising those pairs for which both the English sentence and the French sentence are thirty words or less in length. This is a collection of 1,778,620 translations. In an effort to eliminate some of the typographical errors that abound in the text, a English vocabulary is chosen consisting of all of those words that appear at least twice in English sentences in the data, and as a French vocabulary is chosen consisting of all those words that appear at least twice in French sentences in the data. All other words are replaced with a special unknown English word or unknown French word, according as they appear in an English sentence or a French sentence. In this way an English vocabulary of 42,005 words and a French vocabulary of 58,016 words is obtained. Some typographical errors are quite frequent, for example, momento for memento, and so the vocabularies are not completely free of them. At the same time, some words are truly rare, and in some cases, legitimate words are omitted. Adding e.sub.0 to the English vocabulary brings it to 42, 006 words.

Detailed Description Text - DETX (702):

Eleven iterations of the EM process are performed for this data. The process is initialized by setting each of the 2,437,020, 096 translation probabilities, t(f.vertline.e), to 1/58016. That is, each of the 58,016 words in the French vocabulary is assumed to be equally likely as a, translation for each of the 42,006 words in the English vocabulary. For t(f.vertline.e) to be greater than zero at the maximum likelihood solution for one of the models, f and e must occur together in at least one of the translations in the training data. This is the case for only 25,427,016 pairs, or about one percent of all translation probabilities. On the average, then, each English word appears with about 605 French words.

Detailed Description Text - DETX (703):

Table 6 summarizes the training computation. At each iteration, the probabilities of the various alignments of each translation using one model are computed, and the counts using a second, possibly different model are accumulated. These are referred to in the table as the In model and the Out model, respectively. After each iteration, individual values are retrained only for those translation probabilities that surpass a threshold; the remainder are set to the small value (10.sup.-12). This value is so small that it does not affect the normalization conditions, but is large enough that translation probabilities can be resurrected during later

Detailed Description Text - DETX (704):

iterations. As is apparent from columns 4 and 5, even though the threshold is lowered as iterations progress, fewer and fewer probabilities survive. By the final iteration, only 1,620,287 probabilities survive, an average of about thirty-nine French words for each English word.

Detailed Description Text - DETX (705):

As has been described, when the In model is neither Model 1 nor Model 2, the counts are computed by summing over only some of the possible alignments. Many of these alignments have a probability h smaller than that of the Viterbi alignment. The column headed Alignments in Table 6 shows the average number of alignments for which the probability is within a factor of 25 of the probability of the Viterbi alignment in each iteration. As this number drops, the model concentrates more and more probability onto fewer and fewer alignments so that the Viterbi alignment becomes ever more dominant.

Detailed Descripti n Text - DETX (706):

The last column in the table shows the perplexity of the French text given the English text for the In model of the iteration. The likelihood of the training data is expected to increase with each iteration. This likelihood can be thought of as arising from a product of factors, one for each French word in the training data. There are 28,850,104 French words the 28,850,104.sup.th root of the likelihood is the average factor by which the likelihood is reduced for each additional French word. The reciprocal of this root is the perplexity shown in the table. As the likelihood increases, the perplexity decreases. A steady decrease in perplexity is observed as the iterations progress except when a switch from Model 2 as the In model to Model 3 is made. This sudden jump is not because Model 3 is a poorer model than Model 2, but because Model 3 is deficient: the great majority of its probability is squandered on objects that are not strings of French words. As has been explained, deficiency is not a problem. In the description of Model 1, the P(m.vertline.e) was left unspecified. In quoting perplexities for Models 1 and 2, it is assumed that the length of the French string is Poisson with a mean that is a, linear function of the length of the English string. Specifically, it is assumed that Pr(M=m.vertline.e)=(.lambda.l).sup.m e.sup.-.lambda.l /m!, with .lambda. equal to 1.09.

Detailed Description Text - DETX (709):

Tables 8 through 17 show the translation probabilities and fertilities after the final iteration of training for a number of English words. All and only those probabilities that are greater than 0.01 are shown. Some words, like nodding, in Table 8, do not slip gracefully into French. Thus, there are translations like (Il fait signe que oui .vertline. He is nodding), (Il fait un signe de la tete .vertline. He is nodding), (Il fait un signe de tete affirmatif .vertline. He is nodding), or (Il hoche la tete affirmativement .vertline. He is nodding). As a result, nodding frequently has a, large fertility and spreads its translation probability over a variety of words. In French, what is worth saying is worth saying in many different ways. This is also seen with words like should, in Table 9, which rarely has a fertility greater than one but still produces many different words, among them devrait, devraient, devrions, doit, doivent, devons, and devrais. These are (just a fraction of the many) forms of the French verb devoir. Adjectives fare a little better: national, in Table 10, almost never produces more than one word and confines itself to one of nationale, national, nationaux, and nationales, respectively the feminine, the masculine, the masculine plural, and the feminine plural of the corresponding French adjective. It is clear that the models would benefit from some kind of morphological processing to rein in the lexical exuberance of French.

Detailed Description Text - DETX (711):

together with an article. Thus, farmers typically has a fertility 2 and usually produces either agriculteurs or les. Additional examples are in included in Tables 13 through 17, which show the translation and fertility probabilities for external, answer, oil, former, and not.

Detailed Description Text - DETX (712):

FIGS. 37, 38, and 39 show automatically derived alignments for three translations. In the terminology used above, each alignment is b.sup.> (V(f.vertline.e; 2)). It should be understood that these alignments have been found by a process that involves no explicit knowledge of either French or English. Every fact adduced to support them has been discovered automatically from the 1,778,620 translations that constitute the training data. This data, in turn, is the product of a process the sole linguistic input of which is a set of rules explaining how to find sentence boundaries in the two languages. It may justifiably be argued, therefore, that these alignments are inherent in the Canadian Hansard data itself.

Detailed Description Text - DETX (715):

The final example, in FIG. 39, has several features that bear comment. The second word, Speaker, is connected to the sequence l'Orateur. Like farmers above, it has trained to produce both the word that one naturally thinks of as its translation and the article in front of it. In the Hansard data, Speaker always has fertility 2 and produces equally often l'Orateur and le president. Later in the sentence, starred is connected to the phrase marquees de un asterisque. From an initial situation in which each French word is equally probable as a translation of starred, the system has arrived, through training, at a situation where it is able to connect to just the right string of four words. Near the end of the sentence, give is connected to donnerai, the first person singular future of donner, which means to give. It might be better if both will and give were connected to donnerai, but by limiting notions to no more than one word, Model 5 precludes this possibility. Finally, the last twelve words of the English sentence, I now have the answer and will give it to the House, clearly correspond to the last seven words of the French sentence, je donnerai la repose ala Chambre, but, literally, the French is I will give will answer to the

House. There is nothing about now, have, and, or it, and each of these words has fertility 0. Translations that are as far as this from the literal are rather more the rule than the exception in the training data.

Detailed Descripti n Text - DETX (720):

It has been argued above that neither spiritual nor actual deficiency poses a serious problem, but this is not entirely true. Let w(e) be the sum of P(f.vertline.e) over well-formed French strings and let i(e) be the sum over ill-formed French strings. In a deficient model, w(e)+i(e)<1. In this case, the remainder of the probability is concentrated on the event failure and so w(e)+i(e)+P(failure.vertline.e)=1. Clearly, a model is deficient precisely when P(failure.vertline.e)>0. If P(failure.vertline.e)=0, but i(e)>0, then the model is spiritually deficient. If w(e) were independent of e, neither form of deficiency would pose a problem, but because Models 1-5 have no long-term constraints, w(e) decreases exponentially with I. When computing alignments, even this creates no problem because e and f are known. However, for a given f, if the goal is to discover the most probable e, then the product P(e) P(f.vertline.e) is too small for long English strings as compared with short ones. As a result, short English strings are improperly favored over longer English strings. This tendency is counteracted in part by the following modification:

Detailed Description Text - DETX (724):

Models 1 through 5 all assign non-zero probability only to alignments with notions containing no more than one word each. Except in Models 4 and 5, the concept of a notion plays little role in the development. Even in these models, notions are determined implicitly by the fertilities of the words in the alignment: words for which the fertility is greater than zero make up one-word notions; those for which it is zero do not. It is not hard to give a method for extending the generative process upon which Models 3, 4, and 5 are based to encompass multi-word notions. This method comprises the following enhancements:

Detailed Description Text - DETX (735):

.epsilon.(m.vertline.l) string length probabilities

Detailed Description Text - DETX (736):

t(f.vertline.e) translation probabilities

Detailed Description Text - DETX (751):

.epsilon.E(m.vertline.l) string length probabilities

Detailed Description Text - DETX (752):

t(f.vertline.e) translation probabilities

Detailed Description Text - DETX (753):

a(i.vertline.j, l, m,) alignment probabilities

Detailed Description Text - DETX (756):

Assumptions ##EQU60## This model is not deficient. Model 1 is the special case of this model in which the alignment probabilities are uniform: a(i.vertline.j, l, m)=(l+1).sup.-1 for all i.

Detailed Descripti n Text - DETX (768):

t(f.vertline.e) translation probabilities

Detailed Description Text - DETX (769):

n(.o slashed..vertline.e) fertility probabilities

Detailed Description Text - DETX (770):

p.sub.0, p.sub.1 fertility probabilities for e.sub.0

Detailed Description Text - DETX (771):

d(j.vertline.i, l, m) distortion probabilities

Detailed Description Text - DETX (800):

t(f.vertline.e) translation probabilities

Detailed Description Text - DETX (801):

n(.o slashed..vertline.e) fertility probabilities

Detailed Description Text - DETX (802):

P.sub.0, P.sub.1 fertility probabilities for e.sub.0

Detailed Description Text - DETX (803):

d.sub.1 (.DELTA..sub.j .vertline.A, B) distortion probabilities for movement of the first word of a tablet

Detailed Description Text - DETX (804):

d.sub.>1(.DELTA..sub.j .vertline.B) distortion probabilities for movement of other words of a tablet
Here .DELTA..sub.j is an integer; A is an English class; and B is a French class.

Detailed Description Text - DETX (814):

(f.vertline.e) translation probabilities

Detailed Description Text - DETX (815):

n(.o slashed..vertline.e) fertility probabilities

Detailed Description Text - DETX (816):

P.sub.0, P.sub.1 fertility probabilities for e.sub.

Detailed Description Text - DETX (817):

d.sub.L (.DELTA..sub.j .vertline.B, v) distortion probabilities for leftward movement of the first word of a tablet

Detailed Description Text - DETX (818):

d.sub.R (.DELTA..sub.j .vertline.B, v) distortion probabilities for rightward movement of the first word of a tablet

Detailed Description Text - DETX (819):

d.sub.LR (l.sub.– or.sub.– r .vertline.B, v, v') distortion probabilities for choosing left or right movement

Detailed Description Text - DETX (820):

d.sub.>1 (.DELTA..sub.j .vertline.B, v) distortion probabilities for movement of other words of a tablet

Detailed Description Text - DETX (824):

N.B. In the previous section, simplified embodiment of this model in which the probabilities d.sub.LR do not appear is described.

Detailed Description Text - DETX (860):

4201. For the word being considered, finding a good question for each of a plurality of informant sites. These informant sites are obtained from a table 4207 stored in memory of possible informant sites. Possible sites include but are not limited to, the nouns to the right and left, the verbs to the right and left, the words to the right and left, the words two positions to the right or left, etc. A method of finding a good question is described below. This method makes use of a table 4205 stored in memory probabilities derived from Viterbi alignments. These probabilities a-re also discussed below.

Detailed Description Text - DETX (874):

The purpose of source and target transduction is to facilitate tlhe task of the statistical translation. This will be accomplished if the probability distribution Pr (f', e') is easier to model then the original distribution Pr(f, e). In practice this means that e' and f' should encode global linguistic facts about e and f in a local form.

Detailed Description Text - DETX (881):

The probability models. A translation model such as one of the models in Sections 8-10 is used for both P'(F'.vertline.E') and for P(F.vertline.E). A trigram language model such as that discussed in Section 6 is used for both P(E) and P'(E').

Detailed Description Text - DETX (883):

The probability P(f.vertline.e) computed by the translation model requires a sum over alignments as discussed in detail in Sections 8-10. This sum is often too expensive to compute directly since the number of alignments increases exponentially with sentence length. In the mathematical considerations of this Section, this sum will be approximated by the single term corresponding to the alignment, V(f.vertline.e), with greatest probability. This is the Vitcrli approxirnmatio7i already discussed in Sections 8-10 and V(f.vertline.e) is the Viferbi alhg77ment.

Detailed Description Text - DETX (884):

Let c(f.vertline.e) be the expected number of times that e is aligned with f in the Viterbi alignment of a pair of sentences drawn at random from a large corpus of training data,. Let c(.o slashed..vertline.e) be the expected number of times that .epsilon. is aligned wvith .o slashed.words. Then ##EQU87## where c(f.vertline.e; V) is the number of times that e is aligned with f in the alignment A, and c(.o slashed..vertline.e; V) is the number of times that e generates .o slashed. target words in A. The counts above are also expressible as averages with respect to the model: ##EQU88## Probability distributions p(e, f) and p(.o slashed., e) a-re obtained by normalizing the counts c(f.vertline.e) and c(.o slashed..vertline.e): ##EQU89##

Detailed Descripti n Text - DETX (885):

These are the probabilities that are stored in a table of probabilities 4205.) The conditional distributions p (f.vertline.e) and p(.o slashed..vertline.e) are the Viterbi approximation estimates for the parameters of the model. The marginals satisfy ##EQU92## where u(e) and u(f) are the unigram distributions of e and f and .o slashed.(e) =.sigma..sub..o slashed. p(.o slashed..vertline.e) .o slashed. is the average number of source words aligned with

e. These formulae reflect the fact that in any alignment each source word is aligned with exactly one target word.

Detailed Descripti n Text - DETX (889):

where mn is the average length of the source sentences in the training data., and H(F.vertline.E) and H (.PHI..vertline.E) are the conditional entropies for the probability distributions p(f, e) and p(.o slashed., e) ##EQU93## A similar expression for the cross entropy H(E.vertline.F) will now be given. Since

Detailed Description Text - DETX (891):

where H(E) is the cross entropy of P(E) and I(F, E) is the mutual information between f and e for the probability distribution p(f, e).

Detailed Description Text - DETX (893):

Next consider H(E'.vertline.F'). Let e.fwdarw.e' and f.fwdarw.f' be sense labeling transformations of the type discussed above. Assume that these transformations preserve l,iterbi alhgz?77,eizs; that is, if the words e and f are aligned in the iterbi alignment for (f,e), then their sensed versions e' and f' are aligned in the Viterbi alignment for (f',e'). It follows that the word translation probabilities obtained from the Viterbi alignments satisfy p(f,e) =.sigma..sub.f'.epsilon.f p(f', e)=.sigma..sub.e'.epsilon.e p(f, e'), where the sums range over the sensed versions f' of f and the sensed versions e' of e.

Detailed Description Text - DETX (899):

For sensing source sentences, a question about an informant is a function c from the source vocabulary into the set of possible senses. If the informant of f is x, then f is assigned the sense c (x). The function c (x) is chosen to minimize the cross entropy H(E.vertline.F'). From formula. (175), this is equivalent to maximizing the conditional mutual information I(F', E.vertline.f) between E and F' ##EQU96## where p(f, e, x) is the probability distribution obtained by counting the number of times in the Viterbi alignments that e is aligned with f and the value of the informant of f is x, ##EQU97## An exhaustive search for the best c requires a computation that is exponential in the number of values of x and is not practical. In the aforementioned paper entitled "Word-Sense Disambiguation using Statistical Methods" by P. F. Brown, el al., a good c is found using the flip-flop method which is only applicable if the number of senses is restricted to two.

Detailed Description Text - DETX (902):

The method is based on the fact that, up to a constant independent of c, the mutual information I(F', E.vertline.f) can be expressed as an infimuin over conditional probability distributions q(E.vertline.c), ##EQU98## The best value of the information is thus an infimum over both the choice for c and the choice for the q. This suggests the iterative method, depicted in 4401 for obtaining a good c. This method comprises the steps of:

Detailed Description Text - DETX (910):

The methods of sense-labeling discussed above ask a single question about a single word of context. In other embodiments of the sense labeler, this question is the first question in a decision tree. In still other embodiments, rather than using a single informant site to determine the sense of a word, questions from several different informant sites are combined to determine the sense of a word. In one embodiment, this is done by assuming that the probability of an informant word x.sub.i, at informant site i, given a target word e, is independent of an informant word x.sub.j at a different informant site j given the target word e. Also, in other embodiments, the intermediate source and target structure representations are more sophisticated than word sequences, including, but not limited to, sequences of lexical morphemes, case frame sequences, and parse tree structures.

Detailed Descripti n Text - DETX (913):

A number of researchers have developed methods that align sentences according to the words that they contain. (See for example, Deriving translation data. from bilingual text, by Rt.

Detailed Description Text - DETX (931):

In the Hansard example, the English corpus contains 85,016,286 tokens in 3,510,744 sentences, and the French corpus contains 97,857,452 tokens in 3,690,425 sentences. The a-verage English sentence has 24.2 tokens, while the average French sentence is about 9.5% longer with 26.5 tokens. The left-hand side of FIG. 48 shows the raw data for a portion of the English corpus, and the right-hand side shows the same portion after it was cleaned, tokenized, and divided into sentences. The sentence numbers do not advance regularly because the sample has been edited in order to display a variety of phenomena.

Detailed Description Text - DETX (941):

Given these costs, the standard technique of dynamic programming is used to find the alignment between the major anchors with the least total cost. Dynamic programming is described by R. Bellman in the book titled Dynamic Programming, published by Princeton University Press, Princeton, N. J. in 1957. In theory, the time and space required to find this alignment grow as the product of the lengths of the two sequences to be aligned. In practice, however, by using thresholds and the partial traceback technique described by Brown, Spohrer, Hochschild, and Baker in their paper, Partial Traceback and Dynamic Progra-mming, published in the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, in Paris, France in 1982, the time required can be made linear in the length of the sequences, and the space can be made constant. Even so, the computational demand is severe. In the Hansard example, the two corpora were out of alignment in places by as many as 90, 000 sentences owing to mislabelled or missing files.

Detailed Description Text - DETX (947):

The generation of beads is modelled by the two-state Markov model shown in FIG. 50. The allowed beads are shown in FIG. 51. A single sentence in one corpus is assumed to line up with zero, one, or two sentences in the other corpus. The probabilities of the different cases are assumed to satisfy Pr(e)=Pr(f), Pr(eff)=Pr(eef), and Pr (.paragraph..sub.e)=Pr(.paragraph..sub.f).

Detailed Description Text - DETX (948):

The generation of sentence lengths given beads is mo(lele(d as follows. The probability of an English sentence of length l.sub.e given an e-bead is assumed to be the same as the probability of an English sentence of length l.sub.e in the text as a whole. This probability is denoted by Pr(l.sub.e). Similarly, the probability of a French sentence of length l.sub.f given an f-bead is assumed to equal Pr(l.sub.f). For an ef-bead, the probability of an English sentence of length l.sub.e is assumed to equal Pr(l.sub.e) and the log of the ratio of length of the French sentence to the length of the English sentence is assumed to be normally distributed with mean .mu. and variance .sigma..sup.2. Thus, if .tau.=log(l.sub.f /l.sub.e), then

Detailed Description Text - DETX (949):

with a chosen so that the sum of Pr(l.sub.f .vertline.l.sub.e) over positive values of l.sub.f is equal to unity. For an eef-bead, the English sentence lengths are assumed to be independent with equal marginals Pr(l.sub.e), and the log of the ratio of the length of the French sentence to the sum of the lengths of the English sentences is assumed to be normally distributed with the same mean and variance as for an ef-bead. Finally, for an eff-bead, the probability of an English length l.sub.e is assumed to equal Pr(l.sub.e) and the the log of the ratio of the sum of the lengths of the French sentences to the length of the English sentence is assumed to be normally distributed as before. Then, given the sum of the lengths of the French sentences, the probability of a particular pair of lengths, l.sub.1 and l.sub.2 is assumed to be proportional to Pr(l.sub.f1) Pr(L.sub.f2).

Detailed Description Text - DETX (950):

Together, the model for sequences of beads and the model for sentence lengths given beads define a hidden Markov model for the generation of aligned pairs of sentence lengths. Markov Models are described by L. Baum in the article "An Inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process", appearing in Inequalities in 1972.

Detailed Description Text - DETX (951):

The distributions Pr(l.sub.e) and Pr(l.sub.f) are determined from the relative frequencies of various sentence lengths in the data. For reasonably small lengths, the relative frequency is a reliable estimate of the corresponding probability. For longer lengths, probabilities are determined by fitting the observed frequencies of longer sentences to the tail of a Poisson distribution. The values of the other parameters of the Markov model can be determined by from a large

Detailed Description Text - DETX (958):

By repeating this process many thousands of times, an expected error rate of about 0.9% was estimated for the actual frequency of anchor points in the Hansard data. By varying the parameters of the hidden Markov model, the effect of anchor points and paragraph markers on error rate can be explored. With paragraph markers but no anchor points, the expected error rate is 2.0%, with anchor points but no paragraph markers, the expected error rate is 2.3%, and with neither anchor points nor paragraph markers, the expected error rate is 3.2%. Thus, while anchor points and paragraph markers are important, alignment is still feasible without them. This is promising since it suggests that the method is applicable to corpora for which frequent anchor points are not available.

Detailed Description Text - DETX (969):

Word-by-word alignments obtained in this way offer a valuable resource for work in bilingual lexicography and machine translation. For example, a method of cross-lingual sense labeling, described in Section 11, and also in the aforementioned paper, "Word Sense Disambiguation using Statistical Methods", uses alignments obtained in this way as data for construction of a statistical sense-labelling module.

Detailed Description Text - DETX (975):

A set of partial hypotheses is initialized in step 5401. A partial hypothesis is comprised of a target structure and an alignment with some subset of the morphemes in the source structure to be translated. The initial set generated by step) 5401 consists of a single partial hypothesis. The partial target structure for this partial hypothesis is just an empty sequence of morphemes. The alignment is the empty alignment in which no morphemes in the source structure to be translated are accounted for.

Detailed Description Text - DETX (976):

The system then enters a loop through steps 5402, 5403, and 5404, in which partial hypotheses are iteratively extended until a test for completion is satisfied in step 5403. At the beginning of this loop, in step 5402, the existing set of partial hypotheses is examined and a subset of these hypotheses is selected to be extended in the steps which comprise the remainder of the loop. In step 5402 the score for each partial hypothesis is compared to a threshold (the method used to compute these thresholds is described below). Those partial hypotheses with scores greater than threshold are then placed on a list of partial hypotheses to be extended in step 5404. Each partial hypothesis that is extended in step 5404 contains an alignment which accounts for a subset of the morphemes in the source sentence. The remainder of the morphemes must still be accounted for. Each extension of an hypothesis in step 5404 accounts for one additional morpheme. Typically, there are many tens or hundreds of extensions considered for each partial hypothesis to be extended. For each extension a new score is computed. This score contains a contribution from the language model as well as a contribution from the translation model. The language model score is a measure of the plausibility a priori of the target structure associated with the extension. The translation model score is a measure of the plausibility of the partial alignment associated with the extension. A partial hypothesis is considered to be a full hypothesis when it accounts for the entire source structure to be translated. A full hypothesis contains an alignment in which every morpheme in the source structure is aligned with a morpheme in the hypothesized target structure. The iterative process of extending partial hypotheses terminates when step 5402 produces an empty list of hypotheses to be extended. A test for this situation is made on step 5403.

Detailed Description Text - DETX (984):

Here, * is a symbol which denotes a, sequence boundary, and the factor l(the.vertline.*,*) is the trigram language model parameter that serves as an estimate of the probability that the English morpheme the occurs at the beginning of a sentence. The factor n(1.vertline.the) is the translation model parameter that is an estimate of the probability that the English morpheme the has fertility 1, in other words, that the English morpheme the is aligned with only a single French morpheme. The factor t(la.vertline.the) is the translation model parameter that serves as an estimate of the lexial probability that the English morpheme the translates to the French morpheme la. Finally, the factor d(1.vertline.1) is the translation model parameter that serves as an estimate of the distortion probability that a French morpheme will be placed in position 1 of the French structure given that it is aligned with an English morpheme that is in position 1 of the English structure. In the second example in FIG. 57, the English morpheme mother is hypothesized as accounting for the French morpheme mere. The score for this partial hypothesis is

Detailed Description Text - DETX (985):

Here, the final factor d(7.vertline.1) serves as an estimate of the distortion probability that a French morpheme, such as mere, will be place in the 7th position in a source sequence given that it is aligned with an English morpheme such as mother which is in the 1st position in an hypothesized target sequence.

Detailed Description Text - DETX (986):

Now, suppose the partial hypothesis in FIG. 56 is to be extended on some other invocation of step 5404. A common translation of the pair of French morphemes jeune fille is the English morpheme girl. However, since in a preferred embodiment a, partial hypothesis is extended to account for only a single French morpheme at a time, it is not possible to account for both jeune and fille with a single extension. Rather the system first accounts for one of the morphemes, and then on another round of extensions, accounts for the other. This can be done in two ways, either by accounting first for jeune or by accounting first for fille. FIG. 58 depicts the extension that accounts first for fille. The + symbol in FIG. 58 after the the English morpheme girl denotes the fact that in these extensions girl is to be aligned with more French morphemes than it is currently aligned with, in this case, at least two. A morpheme so marked is referred to as open. A morpheme that is not open is said to be closed. A partial hypothesis which contains an open target morpheme is referred to as open, or as an open partial hypothesis. A partial hypothesis which is not open is referred to as closed, or as a closed partial hypothesis. An extension is referred to as either open or closed according to whether or not the resultant partial hypothesis is open or closed. In a preferred embodiment, only the last morpheme in a partial hypothesis can be designated open. The score for the extension in FIG. 58 is ##EQU101## Here, the factor l(girl.vertline.*, the) is the language model parameter that serves as an estimate of the probability with which the English morpheme girl is the second morpheme in a source structure in which the first morpheme is the. The next factor of 2 is the combinatorial factor that is discussed in the section entitled Translation Models and Parameter Estimation. It is factored in, in this case, because the open English morpheme girl is to be aligned with at least two French morphemes. The factor n (i.vertline.girl) is the translation model parameter that serves as an estimate of the probability that the English morpheme girl will be aligned with exactly i French morphemes, and the sum of these parameters for i between 2 and 25 is an estimate of the probability that girl will be aligned with at least 2 morphemes. It is assumed that the probability that an English morpheme will be aligned with more than 25 French morphemes is 0. Note that in a preferred embodiment of the present invention, this sum can be precomputed and stored in memory as a separate parameter. The factor t (fille .vertline. girl) is the translation model parameter that serves as an estimate of the lexical probability that one of the French morphemes aligned with the English morpheme girl will be the French morpheme fille. Finally, the factor d(3.vertline.2) is the translation model parameter that serves as an estimate of the distortion probability that a French morpheme will be placed in position 3 of the French structure given that it is aligned with an English morpheme which is in position 2 of the English structure. This extension score in Equation 184 is multiplied by the score in Equation 182 for the partial hypothesis which is being extended to yield a new score for the partial hypothesis in FIG. 56 of ##EQU102## Consider now an extension to the partial hypothesis in FIG. 58. If a partial hypothesis that is to be extended contains an open morpheme, then, in a preferred embodiment, that hypothesis can only be extended by aligning another morpheme from the source structure with that open morpheme. When such an extension is made, there are two possibilities: 1) the open morpheme is kept open in the extended partial hypothesis, indicating that more source morphemes are to be aligned with that open target morpheme, or 2) the open morpheme is closed indicating that no more source morphemes are to be aligned with that target morpheme. These two cases are illustrated in FIGS. 59 and 60.

Detailed Description Text - DETX (987):

In FIG. 59, an extension is made of the partial alignment in FIG. 58 by aligning the additional French morpheme jeune with the English morpheme girl. In this example the English morpheme girl is then closed in the resultant partial hypothesis. The extension score for this example is ##EQU103## Here, the first quotient adjusts the fertility score for the partial hypothesis by dividing out the estimate of the probability, that girl is aligned with at least two French morphemes and by multiplying in an estimate of the probability that girl is aligned with exactly two French morphemes. As in the other examples, the second and third factors are estimates of the lexical and distortion probabilities associated with this extension.

Detailed Description Text - DETX (991):

Here, the first two factors are the trigram language model estimates of the probabilities with which up follows girl to–wake, and with which her follows to–wake up, respectively. The third factor is the fertility parameter that serves as an estimate of the probability that up is aligned with no source morphemes. The fourth, fifth, and sixth factors are the appropriate fertility, lexical, and distortion parameters associated with the target morpheme her in this partial alignment.

Detailed Description Text - DETX (992):

FIG. 63 shows a similar extension by up her. The difference with the extension in FIG. 62 is that in FIG. 63, the source morpheme her is open. The score for this extension is ##EQU105## The score for this extension differs from the score in Equation 188 in that the fertility parameter n(1.vertline.her) is replaced by the combinatorial factor 2 and the slim of fertility parameters which provides an estimate of the probability that her will be aligned with at least two source morphemes.

Detailed Description Text - DETX (993):

A remaining type of extension is where a partial hypothesis is extended by an additional connection which aligns a source morpheme with the null target morpheme. The score for this type of extension is similar to those described above. No language model score is factored in, and scores from the translation model are factored in, in accordance with the probabilities associated with the null word as described in the section entitled Translation Models and Parameter Estimation.

Detailed Description Text - DETX (995):

Throughout the hypothesis search process, partial hypotheses are maintained in a set of priority queues. In theory, there is a single priority queue for each subset of positions in the source structure. So, for example, for the source structure oui, oui, three positions: oui is in position 1; a comma is in position 2; and oui is in position 3, and there are therefore 2.sup.3 subsets of positions: [], [1], [2], [3], [1,2], [1,3], [2,3], and [1,2,3]. In practice, these priority queues are initialized only on demand, and many less than the full number of queues possible are used in the hypothesis search. In a preferred embodiment, each partial hypothesis is comprised of a sequence of target morphemes, and these morphemes are aligned with a subset of source morphemes. Corresponding to that subset of source morphemes is a priority queue in which the partial hypothesis is stored. The partial hypotheses within a queue are prioritized according to the scores associated with those hypotheses. In certain preferred embodiments the priority queues are limited in size and only the 1000 hypothesis with the best scores are maintained.

Detailed Description Text - DETX (996):

The set of all subsets of a set of source structure positions can be arranged in a subset lattice. For example, the subset lattice for the set of all sets of the set [1, 2, 3] is shown in FIG. 64. In a subset lattice, a parent of a set S is any which contains one less element than S, and which is also a subset of S. In FIG. 64 arrows have been drawn from each set in the subset lattice to each of its parents. For example, the set [2] is a parent of the set [1,2].

Detail d Description Text - DETX (997):

A subset lattice defines a natural partial ordering on a set of sets. Since the priority queues used in hypothesis search are associated with subsets, a subset lattice also defines a natural partial ordering on the set of priority queues. Thus in FIG. 64, there are two parents of the priority queue associated with the subset of source structure positions [1,3]. These two parents are the priority queues associated with the set [1] and [3]. A priority queue Q.sub.1 is said to be an ancestor of another priority Q.sub.2 if 1) Q.sub.1 is not equal to Q.sub.2, and 2) Q.sub.1 is a subset of Q.sub.2. If Q.sub.1 is an ancestor of Q.sub.2, then Q.sub.2 is said to be a descendant of Q.sub.1.

Detailed Description Text - DETX (999):

A priority queue is said to be active if there are partial hypotheses stored in it. An active priority queue is said to be on the frontier if it has no active descendent. The cardinality of a priority queue is equal to the number of elements in the subset with which it is associated. So, for example, the cardinality of the priority queue which is associated with the set [2,3] is 2.

Detailed Description Text - DETX (1000):

The process in step 5402 functions by assigning a threshold to every active priority queue and then places on the list of partial hypotheses to be extended every partial hypothesis on an active priority queue that has an a score that is greater than the threshold for that priority queue. This is depicted in FIG. 66. First, in step 6601 the threshold for every active priority queue is initialized to infinity, in practice, some very large number. Second, in step 6602, thresholds are determined for every priority queue on the frontier.

Detailed Description Text - DETX (1001):

The method by which these thresholds are computed is best described by first describing what the normalizer of a priority queue is. Each priority queue on the frontier corresponds to a set of positions of source morphemes. At each position of these positions is a particular source morpheme. Associated with each morpheme is a number, which in a preferred embodiment is the unigram probability of that source morpheme. These unigram probabilities are estimated by transducing a large body of source text and simply counting the frequency with which the different source morphemes occur. The normalizer for a priority queue is defined to be the product of all the unigram probabilities for the morphemes at the positions in the associated set of source structure positions. For example, the normalizer for the priority queue associated with the set [2,3] for the source structure la jeune fille V--past–3s reveiller sa mere is:

Detailed Description Text - DETX (1002):

For each priority queue Q on the frontier define the normed score of Q to be equal to the score of the partial hypothesis with the greatest score in Q divided by the normalizer for Q. Let Z be equal to the maximum of all normed scores for all priority queues on the frontier. The threshold assigned to a priority queue Q on the frontier is then equal to Z times the normalizer for that priority queue divided by a constant which in a preferred embodiment is 45.

Detailed Description Text - DETX (1003):

After steps 6602, thresholds have been assigned to the priority queues on the frontier, a loop is performed in steps 6604 through 6610. The loop counter i is equal to a different cardinality on each iteration of the loop. The counter i is initialized in step 6604 to the largest cardinality of any active priority queue, in other words, i is initialized to the maximum cardinality of any priority queue on the frontier. On each iteration of the loop the value of i is decreased by 1 until i is equal to 0, at which point the test 6604 is satisfied and the process of selecting partial hypotheses to be extended is terminated.

Detailed Description Text - DETX (1005):

A schematic flow diagram for this processing step 6608 is shown in FIG. 67. The priority queue Q to be processed enters this step at 6701. Steps 6704 through 6707 perform a loop through all partial hypotheses i in the priority queue Q which are greater than the threshold associated with Q. At step 6705 the partial hypothesis i is added to the list of partial hypotheses to be extended. At step 6706 i is used to adjust the thresholds of all active priority queues which are parents of Q. These thresholds are then used when priority queues of lower priority are processed in the loop beginning at step 6604 in FIG. 66.

Detailed Description Text - DETX (1007):

For example, consider the partial hypothesis depicted in FIG. 59. Suppose this is the partial hypothesis i. The two target morphemes the and girl are aligned with the three source morphemes la, jeune, and file which are in source structure positions 1, 2, and 3 respectively. This hypothesis i is therefore in the priority queue corresponding to the set [1,2,3]. The priority queues that are parents of this hypothesis correspond to the sets [1,2], [1,3], and [2,3]. We can use partial hypothesis i to adjust the threshold in each of these priority queues, assuming they are all active, by computing a parent score, score.sub.p from the score score.sub.i associated with the partial hypothesis i. A potentially different parent score is computed for each active parent priority queue. That parent score is then divided by a constant, which in a preferred embodiment is equal to 45. The new threshold for that queue is then set to the minimum of the previous threshold and that parent score.

Detailed Description Text - DETX (1008):

These parent scores are computed by removing from score.sub.i the contributions for each of the source morphemes la, jeune, and fille. For example, to adjust the threshold for the priority queue [2,3], it is necessary to remove the contribution to the score.sub.i associated with the source morpheme in position 1, which is la. This morpheme is the only morpheme aligned with the, so the language model contribution for the must be removed, as well as the translation model contributions associated with la. Therefore: ##EQU106## As another example, to adjust the threshold for the priority queue [1,3], it is necessary to remove the contribution to the score, associated with the source morpheme in position 2, which is jeune. This morpheme is one of two aligned with the target morpheme girl. If the connection between girl and jeune is removed from the partial alignment in FIG. 59, there is still a connection between girl and fille. In other words, girl is still needed in the partial hypothesis to account, for fille. Therefore, no language model component is removed. The parent score in this case is: ##EQU107## Here the first quotient adjust the fertility score, the second adjusts the lexical score and the third adjusts the distortion score.

Detailed Description Text - DETX (1009):

With some thought, it will be clear to one skilled in the art how to generalize from these examples to other situations. In general, a parent score is computed by removing a connection from the partial alignment associated with the partial hypothesis i. Such a connection connects a target morpheme t in the partial target structure associated with the partial hypothesis i and a source morpheme s in a, source structure. If this connection is the only connection to the target morpheme t, then the language model score for t is divided out, otherwise it is left in. The lexical and distortion scores associated with the source morpheme s are always divided out, as is the fertility score associated with the target morpheme t. If n connections remain to the target morpheme t, since n+1 source morphemes are aligned with t in the partial hypothesis i, then the open fertility score serving as an estimate of the probability that at least n+1 source morphemes will be aligned with t is multiplied in.

Detailed Description Text - DETX (1014):

Extensions of partial hypotheses h which are closed are made in lines 17 through 29. First in line 17 the variable s is set to the identity of the source morpheme at position p in the source structure. This morpheme will have a number of possible target translations. In terms of the translation model, this means that there will be a number of target morphemes t for which the lexical parameter t(t.vertline.s) is greater than a certain threshold, which in an embodiment is set equal to 0.001. The list of such target morphemes for a given source morpheme s can be precomputed. In lines 18 through 29 a loop is made through a list of the target morphemes for the source morpheme s. The variable t is set to the target morpheme being processed in the loop. On line 20, an extension is made in which the target morpheme t is appended to the right, end of the partial target structure associated with h and then aligned with the source morpheme at position p, and in which the target morpheme t is open in the resultant partial hypothesis. On line 21, an extension is made in which the target morpheme t is appended to the right end of the partial target structure associated with h and then aligned with the source morpheme at position p, and in which the target morpheme t is closed in the resultant partial hypothesis. On line 22, an extension is made in which the target morpheme t is appended to the null target morpheme in the partial target structure associated with hypothesis h. It is assumed throughout this description of hypothesis search that every partial hypothesis comprises a single null target morpheme.

D tailed Description Text - DETX (1015):

The remaining types of extensions to be performed are those in which the target structure is extended by two morphemes. In such extensions, the source morpheme at position p is aligned with the second of these two target morphemes. On line 23, a procedure is called which creates a list of target morphemes that can be inserted between the last morpheme on the right of the hypothesis h and the hypothesized target morpheme, t. The lists of target morphemes created by this procedure can be precomputed from language model parameters. In particular, suppose t.sub.r is the last morpheme on the right of the partial target structure comprised by the partial hypothesis h. For any target morpheme t.sub.1 the language model provides a score for the three-word sequence t.sub.r t.sub.1 t. In one preferred embodiment this score is equal to an estimate of 1-gram probability for the morpheme t.sub.r, multiplied by an estimate of the probability with 2-gram conditional probability with which t.sub.1 follows t.sub.r, multiplied by an estimate of the 3-gram conditional probability with which t follows the pair t.sub.r t.sub.1. By computing such a score for each target morpheme t.sub.1, the target morphemes can be ordered according to these scores. The list returned by the procedure called on line 23 is comprised of the m best t.sub.1 's which have scores greater than a threshold z. In one embodiment, z is equal to 0.001 and m is equal to 100.

Detailed Description Paragraph Footnote - DEFN (1):

.sup.3 In this equation and in the remainider of this section, the coiivpntion of using uppercase letters (e.g. E) for random variables and lower case lctters (e.g. e) for the values of random variables continues to be used. ##EQU86## The cross entropy measures the average uncertainty that the model has about the target language translation e of a source language sequence f. Here P(e.vertline.f) is the probability according to the model that e is a translation of f and the sum runs over a collection of all S pairs of sentences in a large corpus comprised of pairs of sentences with each pair consisting of a source and target sentence which are translations of one another (See Sections 8-10).

Detailed Description Paragraph Table - DETL (12):

TABLE 8 ______________________________________ Translation and fertility probabilities for nodding. nodding f t(f .vertline. e) .phi. n(.phi. .vertline. e) ______________________________________ signe 0.164 4 0.342 la 0.123 3 0.293 tete 0.097 2 0.167 oui 0.086 1 0.163 fait 0.073 0 0.023 que 0.073 hoche 0.054 hocher 0.048 faire 0.030 me 0.024 approuve 0.019 qui 0.019 un 0.012 faites 0.011

____

Detailed Description Paragraph Table - DETL (13):

TABLE 9 ______________________________________ Translation and fertility probabilities for should. should f t(f .vertline. e) .phi. n(.phi. .vertline. e) ______________________________________ devrait 0.330 1 0.649 devraient 0.123 0 0.336 devions 0.109 2 0.014 faudrait 0.073 faut 0.058 doit 0.058 aurait 0.041 doivent 0.024 devons 0.017 devrais 0.013 ______________________________________

Detailed Description Paragraph Table - DETL (14):

TABLE 10 ______________________________________ Translation and fertility probabilities for national. national f t(f .vertline. e) .phi. n(.phi. .vertline. e) ______________________________________ nationale 0.469 1 0.905 national 0.418 0 0.094 nationaux 0.054 nationales 0.029 ______________________________________

Detailed Description Paragraph Table - DETL (15):

TABLE 11 ______________________________________ Translation and fertility probabilities for the. the f t (f .vertline. e) .phi. n(.phi. .vertline. e) ______________________________________ le 0.497 1 0.746 la 0.207 0 0.254 les 0.155 l' 0.086 ce 0.018 cette 0.011 ______________________________________

Detailed Descripti n Paragraph Table - DETL (16):

TABLE 12 ______________________________________ Translation and fertility probabilities for farmers. farmers f t(f .vertline. e) .phi. n(.phi. .vertline. e) ______________________________________ agriculteurs 0.442 2 0.731 les 0.418 1 0.228 cultivateurs 0.046 0 0.039 producteurs 0.021

Detailed Description Paragraph Table - DETL (17):

TABLE 13 Translation and fertility probabilities for external. external f t(f .vertline. e) .phi. n(.phi. .vertline. e) exterieures 0.944 1 0.967 exterieur 0.015 0 0.028 externe 0.011 exterieurs 0.010

Detailed Description Paragraph Table - DETL (18):

TABLE 14 Translation and fertility probabilities for answer. answer f t(f .vertline. e) .phi. n(.phi. .vertline. e) reponse 0.442 1 0.809 repondre 0.233 2 0.115 repondu 0.041 0 0.074 a 0.038 solution 0.027 repondez 0.021 repondrai 0.016 reponde 0.014 y 0.013 ma 0.010

Detailed Description Paragraph Table - DETL (19):

TABLE 15 Translation and fertility probabilities for oil. oil f t (f .vertline. e) .phi. n(.phi. .vertline. e) petrole 0.558 1 0.760 petrolieeres 0.138 0 0.181 petrolieere 0.109 2 0.057 le 0.054 petrolier 0.030 petroliers 0.024 huile 0.020 Oil 0.013

Detailed Description Paragraph Table - DETL (20):

TABLE 16 Translation and fertility probabilities for former. former f t(f .vertline. e) .phi. n(.phi. .vertline. e) ancien 0.592 1 0.866 anciens 0.092 0 0.074 ex 0.092 2 0.060 precedent 0.054 l' 0.043 ancienne 0.018 ete 0.013

Detailed Description Paragraph Table - DETL (22):

10.1 Summary of Notation E English vocabulary e English word e English sentence l length of e i position in e, i = 0, 1, . . . , l e.sub.i word i of e e.sub.0 the empty notion e.sub.1.sup.i e.sub.1 e.sub.2 . . . e.sub.i F French vocabulary f French word f French sentence m length of f j position in f, j = 1, 2, . . . , m f.sub.j word j of f f.sub.1.sup.j f.sub.1 f.sub.2 . . . f.sub.j a alignment a.sub.j position in e connected to position j of f for alignment a a.sub.1.sup.j a.sub.1 a.sub.2 . . . a.sub.j .phi..sub.i number of positions of f connected to position i of e .phi..sub.1.sup.i .phi..sub.1 .phi..sub.2 . . . .phi..sub.i .tau. tableau - a sequence of tablets, where a tablet is a sequence of French w .tau..sub.i tablet i of .tau. .tau..sub.0.sup.i .tau..sub.0 .tau..sub.1 . . . .tau..sub.i .phi..sub.i length of .tau..sub.i k position within a tablet, k = 1, 2, . . . , .phi..sub.i .tau..sub.ik word k of .tau..sub.i .pi. a permutation of the positions of a tableau .pi..sub.ik position in f for word k of .tau..sub.i for permutation .pi. .pi..sub.i1.sup.k .pi..sub.i1 .pi..sub.i2 . . . .pi..sub.ik V(f .vertline. e) Viterbi alignment for (f .vertline. e) V.sub.i.rarw.j (f .vertline. e) Viterbi alignment for (f .vertline. e) with ij pegged N(a) neighboring alignments of a N.sub.ij (a) neighboring alignments of a with ij pegged b(a) alignment in N(a) with greatest probability b.sup..infin. (a) alignment obtained by applying b repeatedly to a b.sub.i.rarw.j (a) alignment in N.sub.ij (a) with greatest probability b.sub.i.rarw.j.sup..infin. (a) alignment obtained by applying b.sub.i.rarw.j repeatedly to a A(e) class of English word e B(f) class of French word f .DELTA.j displacement of a word in f .nu., .nu..sup.f vacancies in f .rho..sub.i first position in e to the left of i that has non-zero fertility c.sub.i average position in f of the words connected to position i of e [i] position in e of the i.sup.th one word notion .circle-w/dot..sub.i c.sub.[i] P.sub..theta. translation model P with parameter values .theta. C(f,e) empirical distribution of a sample .psi.(P.sub..theta.) log-likelihood objective function R(P.sub..theta. , P.sub..theta.) relative objective function t(f .vertline. e) translation probabilities (All Models) .epsilon.(m .vertline. l) sentence length probabilities (Models 1 and 2) n(.phi. .vertline. e) fertility probabilities (Models 3, 4, and 5) p.sub.0, p.sub.1 fertility probabilities for e.sub.0 (Models 3, 4, and 5) a (i .vertline. j, l, m) alignment probabilities (Model 2) d(j .vertline. i, l, m) distortion probabilities (Model 3) d.sub.1 (.DELTA.j .vertline. A,B) distortion probabilities for the first word of a tablet (Model 4) d.sub.>1 (.DELTA.j .vertline. B) distortion probabilities for the other words of a tablet (Model 4) d.sub.LR (l.sub.-- or.sub.-- r .vertline. B, .nu., .nu.') distortion probabilities for choosing left or right movement (Model 5) d.sub.L (.DELTA.j .vertline. B, .nu.) distortion probabilities for leftward movement of the first word of a tablet (Model 5) d.sub.R (.DELTA.j .vertline. B,

.nu.) distortion probabilities for rightward movement of the first word of a tablet (Model 5) d.sub.>1 (.DELTA.j .vertline. B, .nu.) distortion probabilities for movement of the other words of a tablet (Model 5)

Claims Text - CLTX (3):

generating the target text in the second language based on a combination of a probability of occurrence of an intermediate structure of text associated with a target hypothesis selected from the second language using a target language model, and a probability of occurrence of the source text given the occurrence of said intermediate structure of text associated with said target hypothesis using a target-to-source translation model; and

Claims Text - CLTX (11):

a computer readable program code means for causing a computer to estimate, for each target hypothesis, a first probability of occurrence of said text associated with said target hypothesis using a target language model;

Claims Text - CLTX (12):

a computer readable program code means for causing a computer to estimate, for each target hypothesis, a second probability of occurrence of the source text given the occurrence of said text associated with said target hypothesis using a target-to-source translation model;

Claims Text - CLTX (13):

a computer readable program code means for causing a computer to combine, for each target hypothesis, said first and second probabilities to produce a target hypothesis match score; and

Claims Text - CLTX (25):

a computer readable program code means for causing a computer to determine a fertility model score for each of said partial hypotheses, said partial hypotheses comprising at least one notion and said fertility model score being proportional to a probability that a notion in the target text will generate a specific number of units of linguistic structure in the source text;

Claims Text - CLTX (26):

a computer readable program code means for causing a computer to determine an alignment score for each of said partial hypotheses, said alignment score being proportional to a probability that a unit of linguistic structure in the target text will align with one of zero or more units of linguistic structure in the source text;

Claims Text - CLTX (27):

a computer readable program code means for causing a computer to determine a lexical model score, for each of said partial hypotheses, said lexical model score being proportional to the probability that said units of linguistic structure in the target text of a given partial hypothesis will translate into said units of linguistic structure of source text;

Claims Text - CLTX (28):

a computer readable program code means for causing a computer to determine a distortion model score for each of said partial hypotheses, said distortion model score being proportional to the probability that the source units of linguistic structure will be in a particular position given by a position of the target units of linguistic structure that generated it; and

Claims Text - CLTX (31):

a computer readable program code means for causing a computer to determine a fertility model score for each of said partial hypotheses, said partial hypotheses comprising at least one notion and said fertility model score being proportional to a probability that a notion in the target text will generate a specific number of notion units of linguistic structure in an intermediate structure source text;

Claims Text - CLTX (32):

a computer readable program code means for causing a computer to determine an alignment score for each of said partial hypotheses, said alignment score being proportional to a probability that a unit of linguistic structure in said intermediate structure of target text will align with one of zero or more units of linguistic structure in said intermediate structure of source text;

Claims Text - CLTX (33):

a computer readable program code means for causing a computer to determine a lexical model score, for each of said partial hypotheses, said lexical model score being proportional to a probability that said units of linguistic structure in said intermediate structure of target text of a given partial hypothesis will translate into said units of linguistic structure of said intermediate structure source text;

Claims Text - CLTX (34):

a computer readable program code means for causing a computer to determine a distortion model score for each of said partial hypotheses, said distortion model score being proportional to a probability that a source unit of linguistic structure will be a particular position given the position of the target units of linguistic structure that generated it; and

Claims Text - CLTX (44):

a computer readable program code means for causing a computer to estimate a first score, said first score being proportional to a probability of occurrence of each intermediate target structure of text associated with said target hypotheses using a target structure language model;

Claims Text - CLTX (45):

a computer readable program code means for causing a computer to estimate a second score, said second score being proportional to a probability that said intermediate target structure of text associated with said target hypotheses will translate into said intermediate source structure of text using a target structure-to-source structure translation model;

Claims Text - CLTX (79):

a computer readable program code means for causing a computer to express each of said intermediate target structures as an ordered sequence of units of linguistic structure, and means for multiplying conditional probabilities of said units within an intermediate target structure given an occurrence of previous units of linguistic structure within said intermediate target structure to obtain said first score.

Claims Text - CLTX (80):

21. A computer program product according to claim 20, wherein said conditional probability of each unit of linguistic structure within an intermediate target structure depends only on a fixed number of preceding units within said intermediate target structure.

Claims Text - CLTX (82):

building a parametric translation model to generate a modeled translation probability, comprising the steps of,

Claims Text - CLTX (85):

choosing a first specification of parameters for the translation model so that the modeled translation probability of the source and target training texts is a first unique local maximum value;

Claims Text - CLTX (86):

building a parametric language model to generate a modeled probability, comprising the steps of,

Claims Text - CLTX (88):

choosing a second specification of parameters for the language model so that the modeled probability of the given training text is a second unique local maximum value; and

Claims Text - CLTX (95):

determining a probability associated with each of said sentence alignments.

Claims Text - CLTX (102):

determining a probability associated with each of said unit alignments.

Claims Text - CLTX (104):

determining a sum of probabilities for all alignments of units of linguistic structure for a given pair of source and target sentences.

Claims Text - CLTX (115):

determining, for a given pair of source and target sentences and a completely specified alignment of units of linguistic structure between the sentences, the probability of said alignment.

Claims Text - CLTX (117):

inputting the first text into the lexical model, wherein the lexical model comprises a parametric translation model for generating a first probability and a parametric language model for generating a second probability; and

Claims Text - CLTX (118):

determining, using the lexical model, the second text in the second language that yields a unique local maximum value of a product of the first probability of the parametric translation model and the second probability of the parametric language model.

Other Reference Publication - OREF (8):

F. Jelinek, R. Mercer, "Interpolated Estimated of Markov Source Parameters From Sparse Data", Workshop on Pattern Recognition in Practice, Amsterdam (Netherland), North Holland, May 21-23, 1980.

Other Reference Publication - OREF (11):

L. Baum, "An Inequality and Associated Maximation Technique in Statistical Estimation for Probalistic Functions of Markov Processes", Inequalities, vol. 3, 1972, pp. 1-8.

Other Reference Publicati n - OREF (15):

Catizone, et al: "Deriving Translation Data from Bilingual Texts"; Proceeding of the First International Acquisition Workshop, Detroit, Mich. 1989, pp. 1-6.